# Applications of the Noncentral t–Distribution

Stat 498B Industrial Statistics

*Fritz Scholz*

April 26, 2007

## 1 Introduction.

The noncentral t–distribution is intimately tied to statistical inference procedures for samples from normal populations. For simple random samples from a normal population the usage of the noncentral t–distribution includes basic power calculations, variables acceptance sampling plans (MIL–STD–414) and confidence bounds for percentiles, tail probabilities, statistical process control parameters $C_L$, $C_U$ and $C_{pk}$ and for coefficients of variation.

The purpose of these notes is to describe these applications in some detail, giving sufficient theoretical derivation so that these procedures may easily be extended to more complex normal data structures, that occur, for example, in multiple regression and analysis of variance settings.

We begin by giving a working definition of the noncentral t–distribution, i.e., a definition that ties directly into all the applications. This is demonstrated upfront by exhibiting the basic probabilistic relationship underlying all these applications.

Separate sections deal with each of the applications outlined above. The individual sections contain no references. However, a short list is provided at the end to give an entry into the literature on the noncentral t–distribution.

For many of the computations we use the R functions `qnct` and `del.nct`. They represent the quantile function and the inverse $\delta$-function of the noncentral $t$-distribution. They do not yet exist in the standard distribution of R. These functions and all other R code used here are provided as part of an R work space at the class web site

`http://www.stat.washington.edu/fritz/Stat498B.html`.

The statistical package R is freely available under the terms of the Free Software Foundation's GNU General Public License for various operating systems (Unix, Linux, Windows, MacOS X) at

`http://cran.r-project.org/`.

## 2 Testing for Normality

Since the noncentral t-distribution arises in applications of normal random samples we discuss here briefly how to examine or test whether such samples indeed come from a normal distribution. There are informal ways via QQ-plots and formal goodness-of-fit tests, of which we only discuss the main tests based on the empirical distribution function (EDF), also referred to as EDF goodness-of-fit tests.

### 2.1 QQ-Plots

To construct a QQ-plot we sort the sample $X_1, \ldots, X_n$ in increasing order $X_{(1)} \leq \ldots \leq X_{(n)}$, assigning fractional ranks $p_i \in (0,1)$, $i = 1, \ldots, n$, to these order statistics in one of several possible and popularly used ways:

$$p_i = \frac{i - .5}{n} \qquad \text{or} \qquad p_i = \frac{i}{n+1} \qquad \text{or} \qquad p_i = \frac{i - .375}{n + .25} \; .$$

Then we plot $X_{(i)}$ against the standard normal $p_i$-quantile $z_{p_i} = \texttt{qnorm}(\texttt{p}_\texttt{i})$ for $i = 1, \ldots, n$. We would expect the sample $p_i$-quantile $X_{(i)}$ to be somewhat close to the corresponding population $p_i$-quantile $x_{p_i} = \mu + \sigma z_{p_i}$, at least as the sample size $n$ gets larger. The pattern of the plotted points $(z_{p_i}, X_{(i)})$ should look therefore look approximately linear with intercept $\approx \mu$ and slope $\approx \sigma$. However, judging approximate linearity takes practice.

Daniel and Wood (1971) recognized the importance of practicing this judgment, in particular in relation to the sample size $n$, by including several pages with many such normal sample plots in their book. Nowadays it has become quite easy to gain such experience by generating (standard) normal random samples of size $n$ in R (or S-Plus) by using `x=rnorm(n` and following that with the command `qqnorm(x)` which produces the corresponding QQ-plot. R uses the third choice for $p_i$, presented above. To judge the linearity one can follow this up with the command `qqline(x)` which superimposes a fitted line on the QQ-plot. The line is fitted to the middle half of the data.

Some such QQ-plots are shown in Figures 1-4 for sample sizes $n = 8, 16, 64, 256$. While at sample size $n = 8$ such QQ-plots can exhibit strong nonlinear patterns, this subsides as $n$ gets large. For large $n$ one can still expect some fluctuating behavior in the tails. That is not unusual and should not necessarily be construed as evidence of nonlinearity and thus nonnormality. Intuitively such sample tail fluctuations can be understood by the fact that near the sample extremes the data are not hemmed in quite as strongly as they are in the main part of the sample. When QQ-plots are not clearly linear one should resort to using EDF goodness-of-fit tests to clarify the issue. However, even such tests may then give an ambiguous verdict. On the other hand, one may want to do both routinely, the QQ-plot for visual impression of the data and the routine EDF goodness-of-fit test to avoid sample selection bias, which would invalidate the calculated $p$-values.

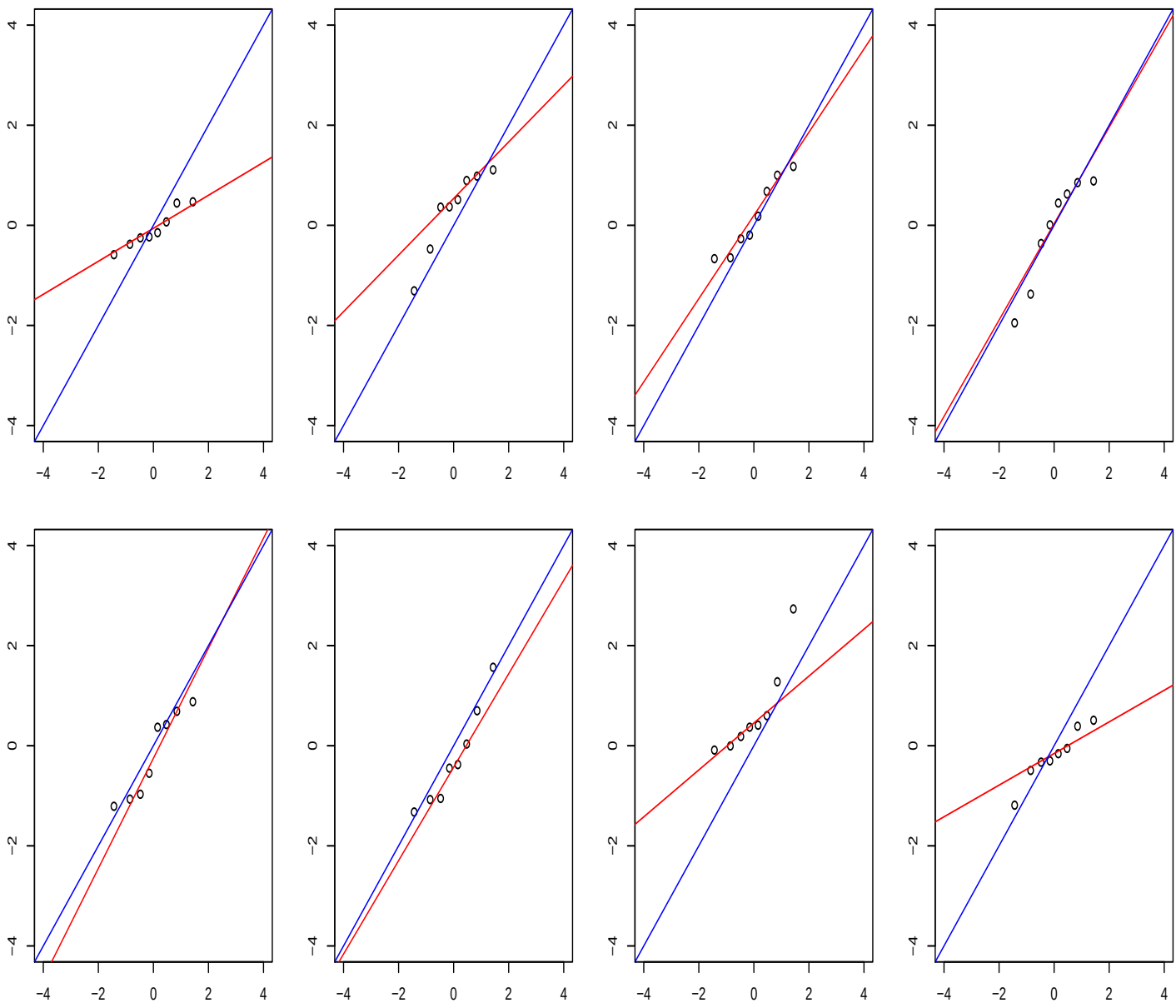



Figure 1: Standard Normal Samples of Size $n = 8$

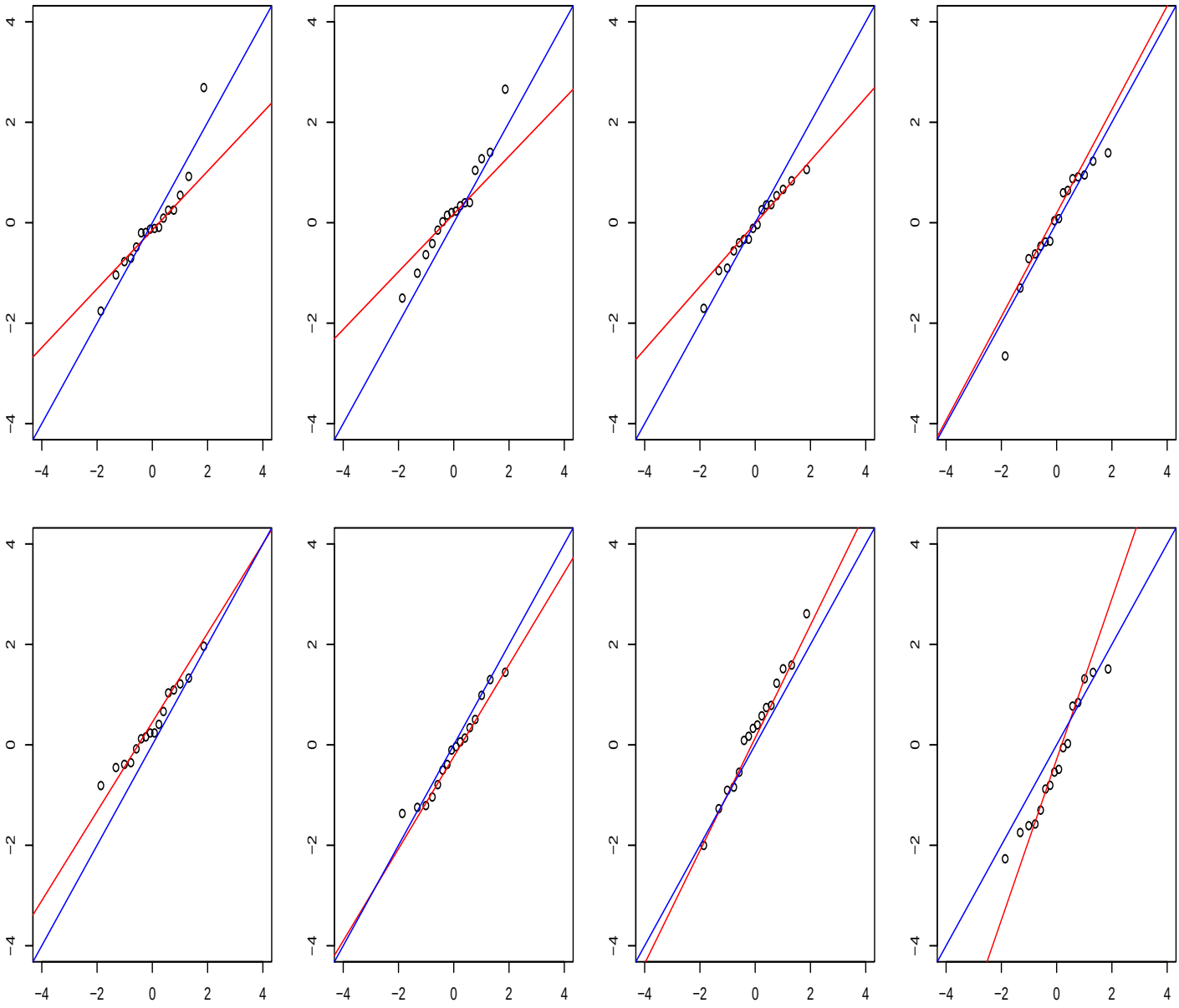



Figure 2: Standard Normal Samples of Size $n = 16$

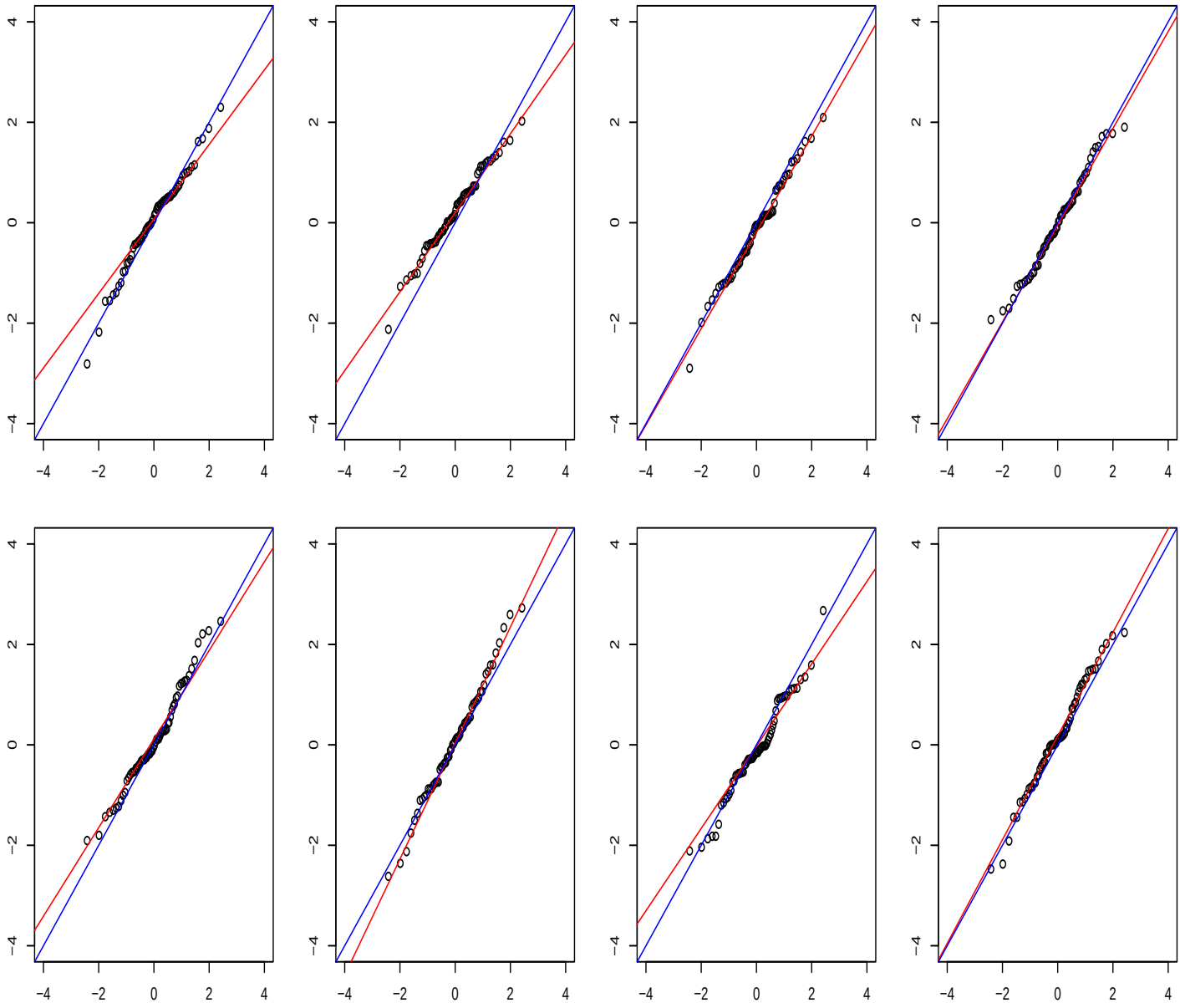



Figure 3: Standard Normal Samples of Size $n = 64$

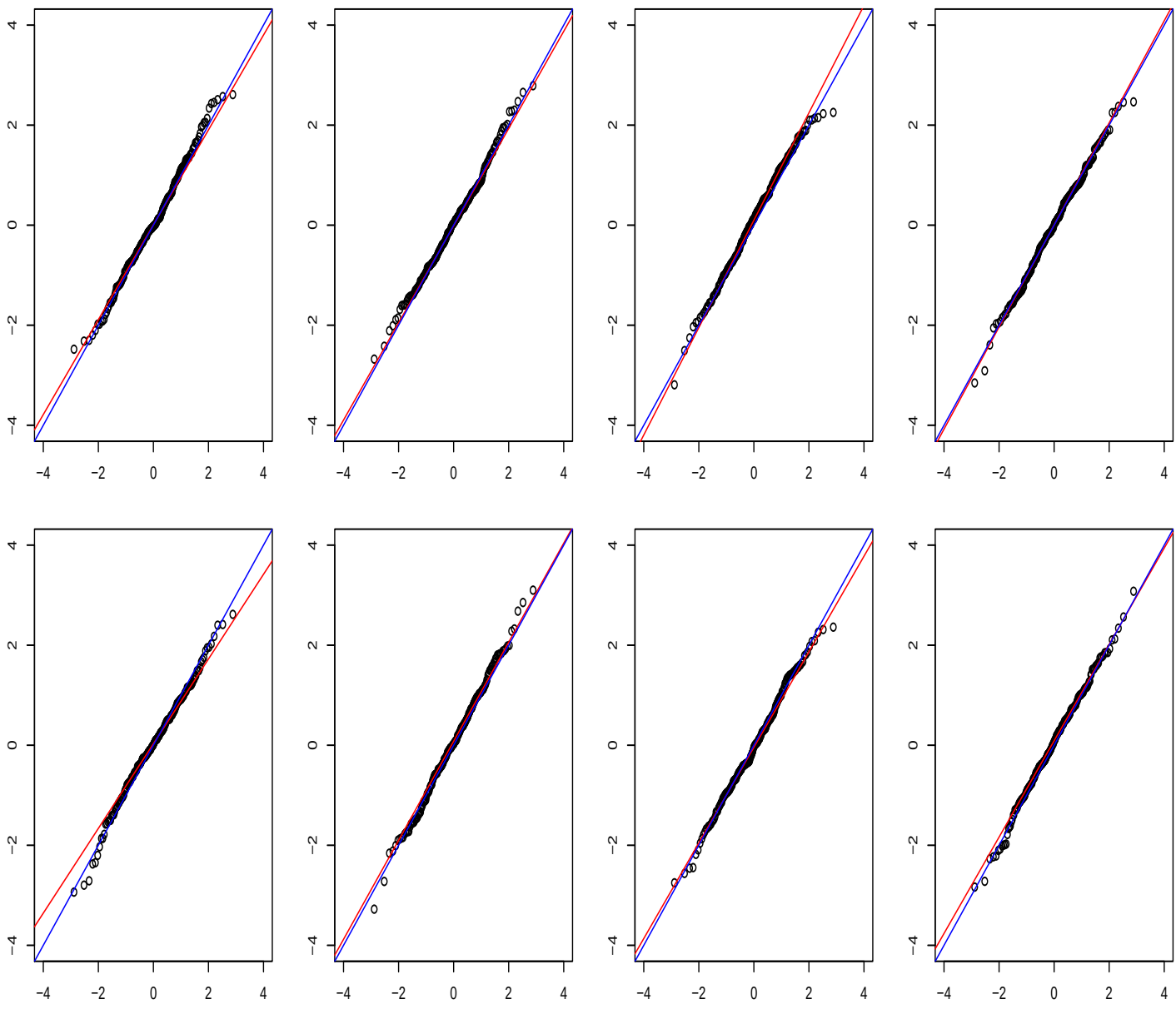



Figure 4: Standard Normal Samples of Size $n = 256$

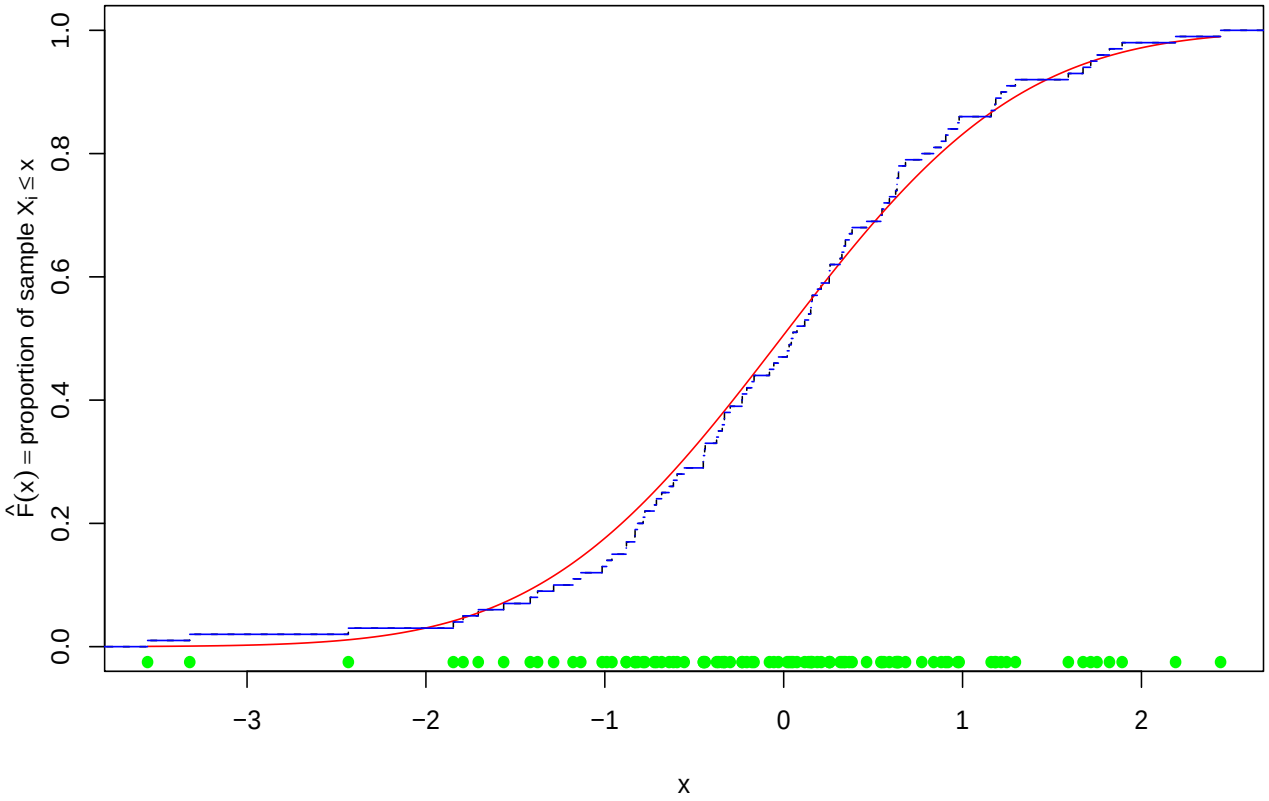



Figure 5: Normal Sample of Size $n = 100$, EDF and True CDF Comparison

### 2.2 EDF Goodness-of-Fit Tests

We can also carry out formal EDF-based tests of fit for normality. Assume that $X_1, \ldots, X_n \sim F$. We wish to test the composite $H_0 : F(x) = \Phi((x - \mu)/\sigma)$ for some $\mu$ and $\sigma$ (composite, because under the hypothesis $\mu$ and $\sigma > 0$ can take on any values).

The empirical distribution function (EDF) is defined as

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} I_{(-\infty,\, x]}(X_i) \quad \text{with} \quad I_{(-\infty,\, x]}(X_i) = 1 \ \text{ or } \ 0 \quad \text{as} \quad X_i \leq x \ \text{ or } \ X_i > x \ ,$$

i.e., $F_n(x)$ is the proportion of sample values $\leq x$. By the Law of Large Numbers (LLN) we have that $F_n(x) \overset{n\to\infty}{\longrightarrow} F(x)$ for all $x$. In fact, since $F(x)$ is continuous this convergence is uniform in $x$. Figure 5 shows such the EDF for a normal sample of size $n = 100$ in relation to the sampled distribution function $F(x)$. The fit looks quite good.

Since we do not know the specific distribution $F(x)$ from which the sample was drawn, even under the hypothesis ($\mu$ and $\sigma$ are unknown), we use its natural estimate $\hat{F}(x) = \Phi((x - \bar{X})/S)$ to compare it with $F_n(x)$ via some discrepancy metric $D(F_n, \hat{F})$.

The following are the three main discrepancy metrics that are in use

- The Kolmogorov-Smirnov metric (measuring maximum local discrepancies)

$$D = \max_x \left\{ \left| \hat{F}_n(x) - \Phi\left( \frac{x - \bar{X}}{S} \right) \right| \right\}$$

- The Cramer-von-Mises metric (measuring cumulative discrepancies)

$$W^2 = \int_{-\infty}^{\infty} \left[ \hat{F}_n(x) - \Phi\left( \frac{x - \bar{X}}{S} \right) \right]^2 \frac{1}{S} \varphi\left( \frac{x - \bar{X}}{S} \right) dx \quad \text{with} \quad \varphi(x) = \Phi'(x)$$

- The Anderson-Darling metric (cumulative discrepancies with sensitivity to tail behavior)

$$A^2 = \int_{-\infty}^{\infty} \frac{\left[ \hat{F}_n(x) - \Phi\left( \frac{x - \bar{X}}{S} \right) \right]^2}{\Phi\left( \frac{x - \bar{X}}{S} \right) \left[ 1 - \Phi\left( \frac{x - \bar{X}}{S} \right) \right]} \frac{1}{S} \varphi\left( \frac{x - \bar{X}}{S} \right) dx$$

The form of these metrics are mainly given to better understand their nature. For computational purposes one uses the following equivalent expressions, all of which employ simple summations over $i = 1, \ldots, n$.

- The Kolmogorov-Smirnov metric

$$D = \max \left[ \max \left\{ \frac{i}{n} - \Phi\left( \frac{X_{(i)} - \bar{X}}{S} \right) \right\}, \max \left\{ \Phi\left( \frac{X_{(i)} - \bar{X}}{S} \right) - \frac{i - 1}{n} \right\} \right]$$

- The Cramer-von-Mises metric

$$W^2 = \sum_{i=1}^{n} \left\{ \Phi\left( \frac{X_{(i)} - \bar{X}}{S} \right) - \frac{2i - 1}{2n} \right\}^2 + \frac{1}{12n}$$

- The Anderson-Darling metric

$$A^2 = -n - \frac{1}{n} \sum_{i=1}^{n} \left[ (2i - 1) \log\left( \Phi\left( \frac{X_{(i)} - \bar{X}}{S} \right) \right) \right.$$
$$\left. + (2n + 1 - 2i) \log\left( 1 - \Phi\left( \frac{X_{(i)} - \bar{X}}{S} \right) \right) \right]$$

It turns out that the distributions of the all three discrepancy metrics given above are independent of $\mu$ and $\sigma$. However, these distributions depend on $n$. For each discrepancy metric and each $n$ there is thus only one such distribution under the composite hypothesis $H_0$. In principle it can be generated by simulation for any fixed $n$, by simulating $N_{\text{sim}} = 10000$ such samples of size $n$ and computing $D(F_n, \hat{F})$ for each such sample. That way we would get a close approximation to this null distribution. In fact, the tabled values of this null distribution are based on a combination of simulations and large sample asymptotic results. Using this null distribution for $D(F_n, \hat{F})$ we reject $H_0$ at significance value $\alpha$ whenever $D(F_n, \hat{F}) > d_{1-\alpha,n}$. Here $d_{p,n}$ is the $p$-quantile of the $D(F_n, \hat{F})$ distribution for sample size $n$. Such $p$-quantiles are given in D'Agostino and M.A. Stephens (1986). In R these tests can be enabled by installing the package `nortest_1.0.zip` from the class web site to the directory that houses your R work space. Under the R `Packages` menu item choose "Install package(s) from local zip files." and proceed from there. This installation is done only once on your computer for the installed version of R. After this installation you need to invoke `library(nortest)` in any R session during which you wish to use the functions in the package `nortest`. These functions are `lillie.test`, `cvm.test` and `ad.test` and you get documentation on them by placing a `?` in front of the respective function names, e.g., `?lillie.test`. Here `?lillie.test` alludes to the Lilliefors (Kolmogorov-Smirnov) test with special steps taken for the $p$-value computation.

```
> lillie.test(rnorm(7))
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  rnorm(7)
D = 0.287, p-value = 0.08424

> lillie.test(runif(137))
        Lilliefors (Kolmogorov-Smirnov) normality test
data:  runif(137)
D = 0.0877, p-value = 0.01169

> ad.test(rnorm(10))
        Anderson-Darling normality test
data:  rnorm(10)
A = 0.4216, p-value = 0.2572
> ad.test(runif(30))
        Anderson-Darling normality test
data:  runif(30)
A = 0.8551, p-value = 0.02452
```

## 3 Definition of the Noncentral t–Distribution

If $Z$ and $V$ are (statistically) independent standard normal and chi–square random variables respectively, the latter with $f$ degrees of freedom, then the ratio

$$T_{f,\,\delta} = \frac{Z+\delta}{\sqrt{V/f}}$$

is said to have a noncentral t–distribution with $f$ degrees of freedom and noncentrality parameter $\delta$. Although $f \geq 1$ originally was intended to be an integer closely linked to sample size, it is occasionally useful to extend its definition to any real $f > 0$. The noncentrality parameter $\delta$ may also be any real number. The cumulative distribution function (cdf) of $T_{f,\,\delta}$ is denoted by $G_{f,\,\delta}(t) = P(T_{f,\,\delta} \leq t)$. If $\delta = 0$, then the noncentral t–distribution reduces to the usual central or Student t–distribution. $G_{f,\,\delta}(t)$ increases from 0 to 1 as $t$ increases from $-\infty$ to $+\infty$ and it decreases from 1 to 0 as $\delta$ increases from $-\infty$ to $+\infty$. While the former is a standard property of any cdf, the latter becomes equally obvious when rewriting $G_{f,\,\delta}(t)$ as follows

$$G_{f,\,\delta}(t) = P\left(\frac{Z+\delta}{\sqrt{V/f}} \leq t\right) = P\left(Z - t\sqrt{V/f} \leq -\delta\,.\right)$$

This monotonicity w.r.t. $\delta$ is illustrated in Figure 6 where the cdfs clearly move to the right as $\delta$ increases or they are vertically ordered, which expresses the above monotonicity with respect to $\delta$ at any fixed value $t$. There appears to be no such simple monotonicity relationship with regard to the parameter $f$. This is illustrated in Figure 7, where the cdfs cross each other, although not at the same point, even though the plot may give that appearance. As $f$ gets very large the distribution of $T_{f,\,\delta}$ approximates the normal distribution of $Z+\delta$. This can also be seen in Figure 7 when focussing on the case df = 100. Figures 6 and 7 were produced by `density.plot.delta` and `density.plot.df`, respectively.

In R the cdf $G_{f,\,\delta}(t)$ is evaluated using the function call `pt(t,f,`$\delta$`)` while its density at $t$ is evaluated by `dt(t,f,`$\delta$`)`. R gives a corresponding quantile-function only when $\delta = 0$, i.e., for the central $t$-distribution. We provide such a function `qnct` for any $\delta$ in the referenced R work space. Similarly, the inverse of $G_{f,\,\delta}(t)$ with respect to $\delta$ is useful and is given there as `del.nct`.

Since most of the applications to be treated here concern single samples from a normal population, we will review some of the relevant normal sampling theory. Suppose $X_1,\ \ldots,\ X_n$ is a random sample from a normal population with mean $\mu$ and standard deviation $\sigma$. The sample mean $\bar{X}$ and sample standard deviation $S$ are respectively defined as:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}\,.$$

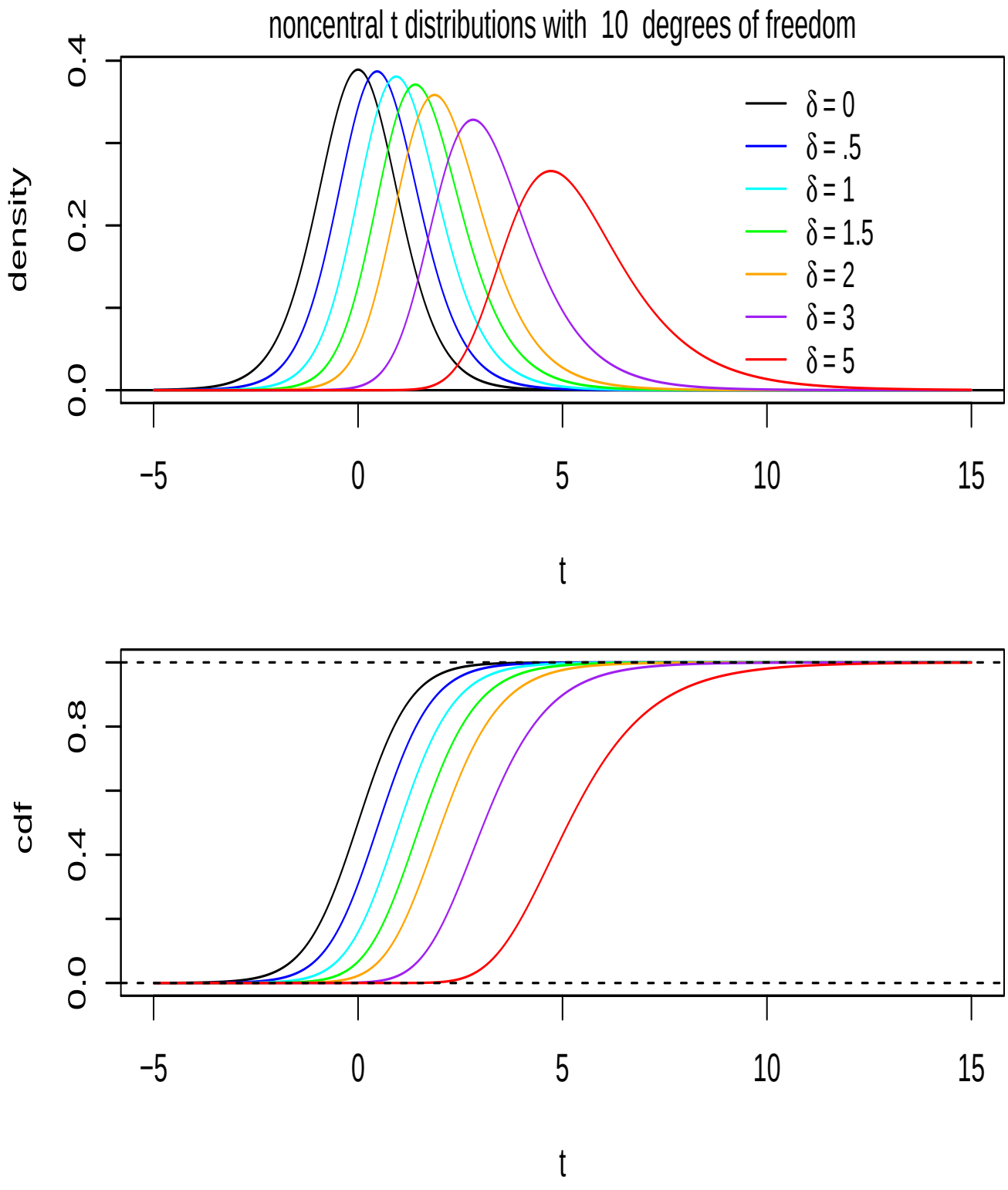



Figure 6: The Effect of $\delta$ on the Noncentral t-Distribution

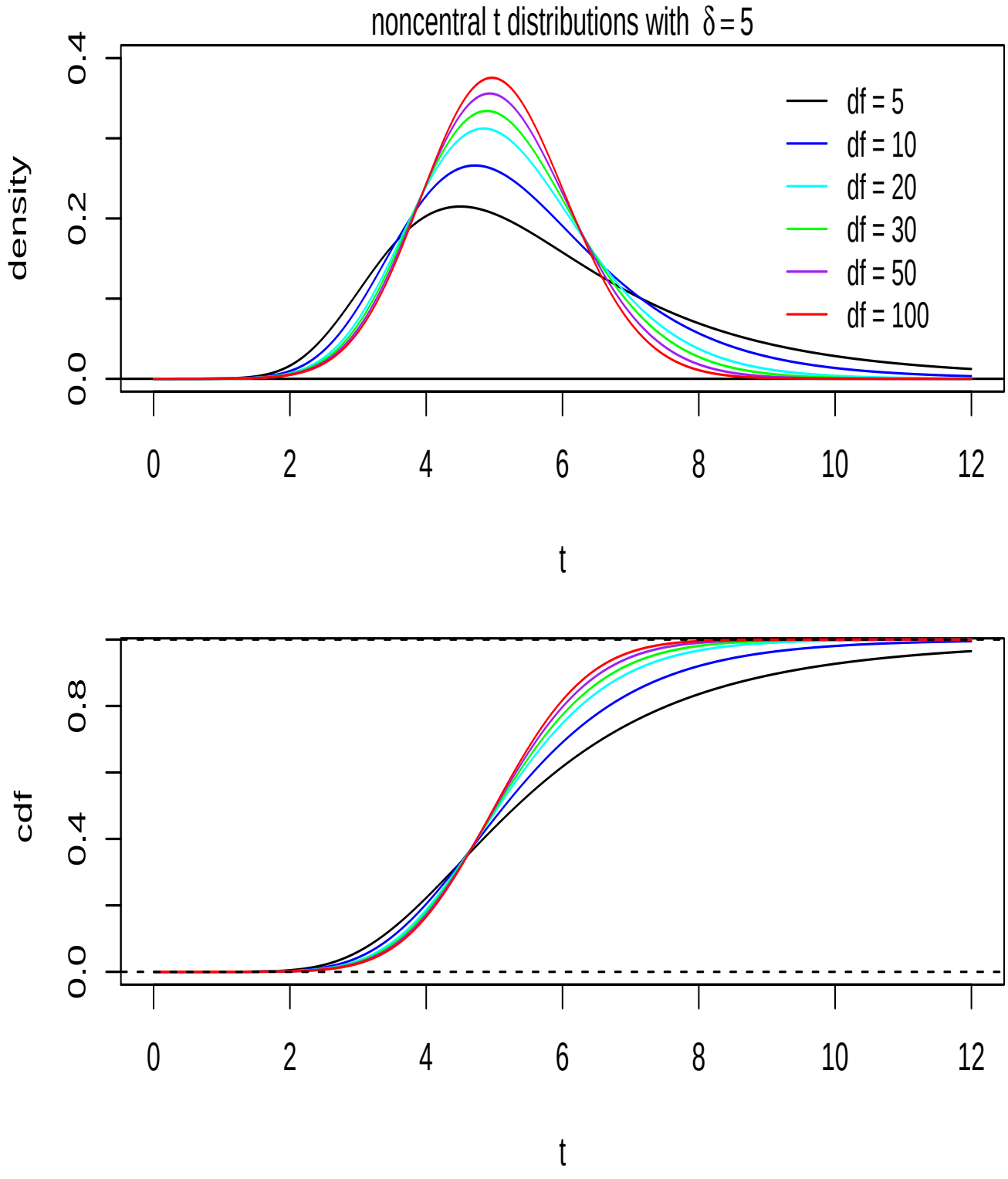



Figure 7: The Effect of df on the Noncentral t-Distribution

The following distributional facts are well known:

- $\bar{X}$ and $S$ are statistically independent;
- $\bar{X}$ is distributed like a normal random variable with mean $\mu$ and standard deviation $\sigma/\sqrt{n}$, or equivalently, $Z = \sqrt{n}(\bar{X} - \mu)/\sigma$ has a standard normal distribution (mean = 0 and standard deviation = 1);
- $V = (n-1)S^2/\sigma^2$ has a chi-square distribution with $f = n - 1$ degrees of freedom and is statistically independent of $Z$.

All one–sample applications involving the noncentral t–distribution can be reduced to calculating the following probability

$$\gamma = P(\bar{X} - aS \leq b) \; . \tag{1}$$

To relate this probability to the noncentral t–distribution note the equivalence of the following three inequalities, which can be established by simple algebraic manipulations:

$$\bar{X} - aS \leq b \iff \frac{\sqrt{n}(\bar{X} - \mu)/\sigma - \sqrt{n}(b - \mu)/\sigma}{S/\sigma} \leq a\sqrt{n} \iff T_{f,\,\delta} \stackrel{\text{def}}{=} \frac{Z + \delta}{\sqrt{V/f}} \leq a\sqrt{n}$$

with $f = n - 1$, $\delta = -\sqrt{n}(b - \mu)/\sigma$, and with $Z$ and $V$ as defined previously in terms of $\bar{X}$ and $S$. Thus

$$\gamma = P(T_{f,\,\delta} \leq a\sqrt{n}) = G_{f,\,\delta}(a\sqrt{n}) \; . \tag{2}$$

Depending on the application, three of the four parameters $n$, $a$, $\delta$ and $\gamma$ are usually given and the fourth needs to be determined either by direct computation of $G_{f,\,\delta}(t)$ or by root solving techniques, using `qnct` or `del.nct`, or by iterative trial and error with $n$.

The following identity is sometimes useful and is based on the fact that $Z$ and $-Z$ have the same distribution:

$$\begin{aligned} G_{f,-\delta}(-t) &= P\left(\frac{Z - \delta}{\sqrt{V/f}} \leq -t\right) = P\left(\frac{-Z + \delta}{\sqrt{V/f}} \geq t\right) \\ &= P\left(\frac{Z + \delta}{\sqrt{V/f}} \geq t\right) = 1 - P\left(\frac{Z + \delta}{\sqrt{V/f}} \leq t\right) = 1 - G_{f,\delta}(t) \end{aligned} \tag{3}$$

## 4 Power of the t–Test

Assuming the normal sampling situation described above, the following testing problem is often encountered. A hypothesis $H : \mu \leq \mu_0$ is tested against the alternative $A : \mu > \mu_0$. Here $\mu_0$ is some specified value. For testing $H$ against $A$ on the basis of the given sample, the intuitive and in many ways optimal procedure is to reject $H$ in favor of $A$ whenever

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \geq t_{n-1}(1-\alpha) \qquad \text{or equivalently when} \qquad \bar{X} - \frac{t_{n-1}(1-\alpha)\, S}{\sqrt{n}} \geq \mu_0 \ .$$

Here $t_{n-1}(1-\alpha)$ is the $1-\alpha$ percentile of the central t–distribution with $n-1$ degrees of freedom. In this form the test has chance $\alpha$ or less of rejecting $H$ when $\mu \leq \mu_0$, i.e., when $H$ is true. As will become clear below, the chance of rejection is $< \alpha$ when $\mu < \mu_0$. Thus $\alpha$ is the maximum chance of rejecting $H$ falsely, i.e., the maximum type I error probability.

An important characteristic of a test is its power function, which is defined as the probability of rejecting $H$ as a function of $(\mu,\ \sigma)$, i.e.,

$$\beta(\mu,\ \sigma) = P_{\mu,\,\sigma}\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \geq t_{n-1}(1-\alpha)\right) \ .$$

The arguments and subscripts $(\mu,\ \sigma)$ indicate that the probability is calculated assuming that the sample $X_1,\ \ldots,\ X_n$ comes from a normal population with mean $\mu$ and standard deviation $\sigma$.

For $\mu > \mu_0$ the value of $1 - \beta(\mu,\ \sigma)$ represents the probability of falsely accepting $H$, i.e., the probability of type II error. The power function can be expressed directly in terms of $G_{f,\,\delta}(t)$ by noting

$$\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma + \sqrt{n}(\mu - \mu_0)/\sigma}{S/\sigma} = \frac{Z + \delta}{\sqrt{V/(n-1)}} \ ,$$

so that

$$\beta(\mu,\ \sigma) = P_{\mu,\,\sigma}\left(\frac{\sqrt{n}(\bar{X} - \mu_0)}{S} \geq t_{n-1}(1-\alpha)\right) = 1 - G_{n-1,\,\delta}(t_{n-1}(1-\alpha)) \ ,$$

where $\delta = \sqrt{n}(\mu - \mu_0)/\sigma = \sqrt{n}\Delta$ with $\Delta = (\mu - \mu_0)/\sigma$.

This power function depends on $\mu$ and $\sigma$ only through the noncentrality parameter $\delta$. It strictly increases from 0 to 1 as $\delta$ increases from $-\infty$ to $\infty$. Thus the maximum rejection probability under $H$ occurs at $\mu = \mu_0$ $(\delta = 0)$, as claimed previously.

We point out that with increasing sample size $n$ the noncentrality parameter $\delta$ can become arbitrarily large. Thus we will reject $H$ for any alternative $\mu > \mu_0$ with probability increasing to 1, no matter how close $\mu$ is to $\mu_0$ and no matter how large $\sigma$ is. Of course one should address the practical significance issue of any difference $\mu - \mu_0$ and weigh that against the cost of a large sample size. In

doing so, the magnitude of $\mu - \mu_0$ would typically be judged in relation to the inherent population variability $\sigma$.

For the purpose of planning the necessary sample size $n$ to achieve a given power $\beta$ (or type II error probability $1-\beta$) for a specific alternative it is not sufficient to specify an alternative $\mu > \mu_0$ since that can lead to a continuum of different $\Delta$ values depending on the magnitude of $\sigma$. Instead one should specify the alternative through the parameter $\Delta$, say $\Delta = \Delta_1 > 0$. Thus one is interested in alternatives for which the mean $\mu$ exceeds the hypothesized value $\mu_0$ by $\Delta_1$ units of $\sigma$, whatever $\sigma$ may be.

A complicating issue is that the power function depends on $n$ not only through $\delta = \sqrt{n}\Delta_1$ but also through $t_{n-1}(1-\alpha)$ and through the degrees of freedom $n-1$ in the cdf $G_{n-1,\,\delta}$. The smallest sample size $n$ that achieves a given power $\beta$ at $\Delta_1$ can be found through iteration, starting with a crude initial guess $\tilde{n} = ((z_\beta - z_\alpha)/\Delta_1)^2$ rounded up to the next integer. Here $z_p$ denotes the $p$-quantile of the standard normal distribution. This crude initial guess is based on treating the noncentral $t$-distribution as a $\mathcal{N}(\delta, 1)$ distribution, which it approaches as $n$ gets large.

The R function `min.sample.size` (available in the previously mentioned R work space) carries out this iterative process and reports the initial $\tilde{n}$ and resulting initial power, in addition to the final $n$ and its achieved power $\geq \beta$. This function also produces the plots in Figures 8.

In a similar fashion one can deal with the dual problem of testing the hypothesis $H' : \mu \geq \mu_0$ against the alternative $A' : \mu < \mu_0$. The modifications, which consist of reversing certain inequalities, e.g., rejecting $H'$ when $\sqrt{n}(\bar{X} - \mu_0)/S \leq t_{n-1}(\alpha)$, are straightforward and are omitted.

For the two–sided problem of testing $H^\star : \mu = \mu_0$ against the alternative $A^\star : \mu \neq \mu_0$ the relevant test rejects $H^\star$ in favor of $A^\star$ whenever

$$\frac{\sqrt{n}|\bar{X} - \mu_0|}{S} \geq t_{n-1}(1-\alpha/2)\ .$$

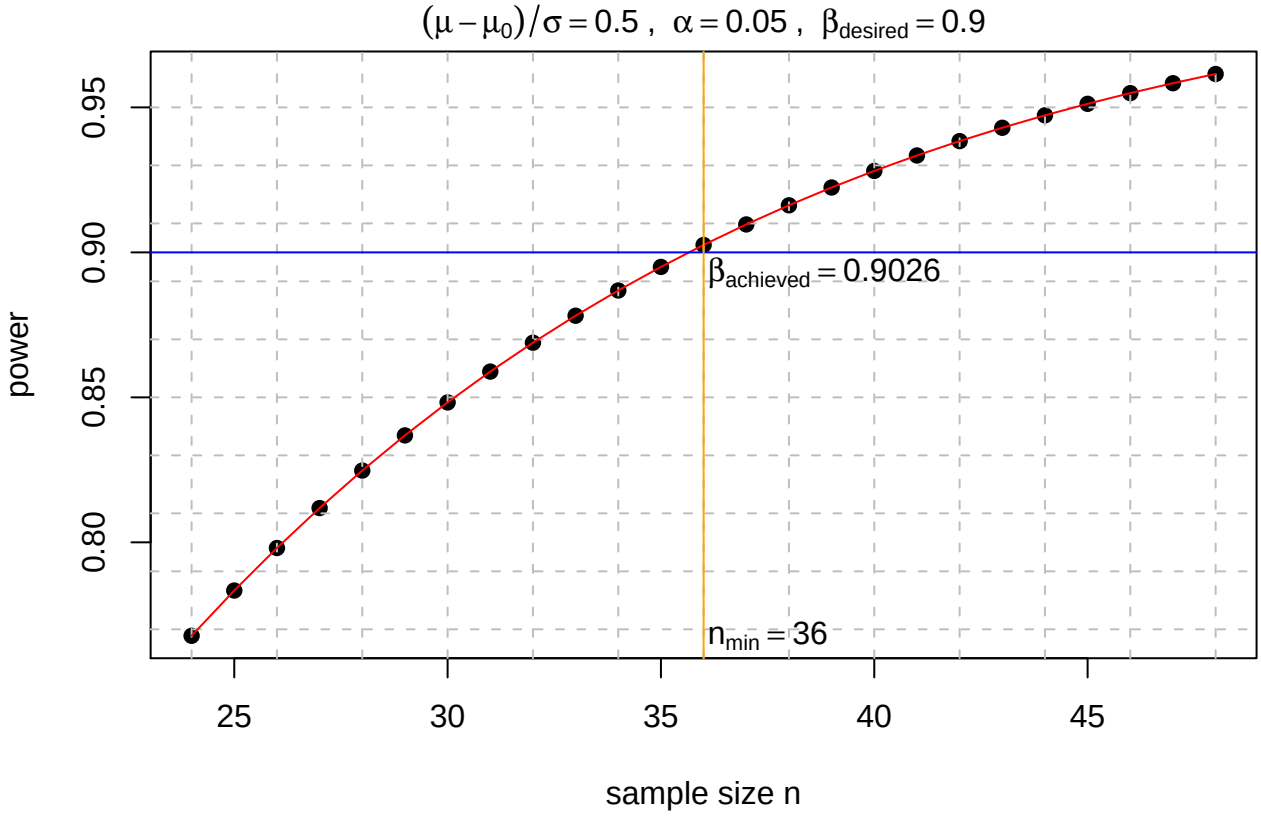
The power function $\beta(\mu, \sigma)$ of this test is calculated along the same lines as before as

$$P_{\mu,\,\sigma}\left(\frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \leq -t_{n-1}(1-\alpha/2) \ \text{ or } \ \frac{\sqrt{n}(\bar{X}-\mu_0)}{S} \geq t_{n-1}(1-\alpha/2)\right)$$

$$= G_{n-1,\,\delta}(-t_{n-1}(1-\alpha/2)) + 1 - G_{n-1,\,\delta}(t_{n-1}(1-\alpha/2)) = \beta^\star(\mu,\ \sigma)\ ,$$

where $\delta = \sqrt{n}(\mu-\mu_0)/\sigma$. It is easy to see that the power function depends on $\mu$ and $\sigma$ only through $|\delta|$ and is strictly increasing in $|\delta|$.

The function `min.sample.size` also determines the minimum required sample sizes for these last two testing scenarios.



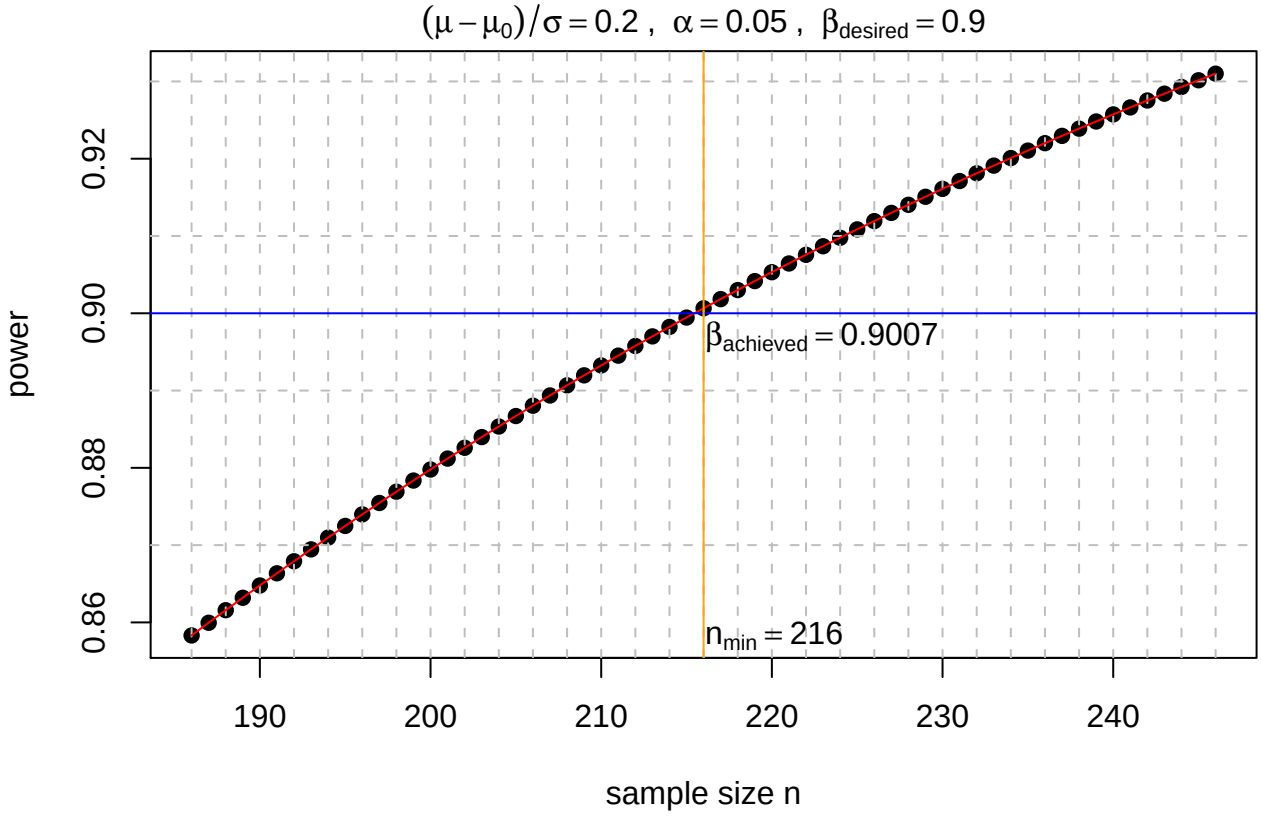




Figure 8: Minimum Sample Size Determination

## 5 Variables Acceptance Sampling Plans

Quality control applications governed by MIL–STD–414 deal with variables acceptance sampling plans (VASP). In a VASP the quality of items in a given sample is measured on a quantitative scale. An item is judged defective when its measured quality exceeds a certain threshold.

The samples are drawn randomly from a population of items. The objective is to make inferences about the proportion of defectives in the population. This leads either to an acceptance or a rejection of the population quality as a whole.

In various applications the term "population" can have different meanings. It represents that collective of items from which the sample is drawn. Thus it could be a shipment, a lot or a batch or any other collective entity. For the purpose of this discussion the term "population" will be used throughout. Ultimately, any batch, lot or shipment is comprised of items that come from a certain process. If that process were to run indefinitely it would produce an infinite population of such items. Thus the sampled items from the batch, lot or shipment could be considered as a sample from that larger conceptual population. If the sample indicates that something is wrong the producer would presumably adjust the process appropriately.

A VASP assumes that measurements (variables) $X_1, \ldots, X_n$ for a random sample of $n$ items from a population are available and that defectiveness for any given sample item $i$ is equivalent to $X_i < L$, where $L$ is some given lower specification limit. In other applications we may call item $i$ defective when $X_i > U$, where $U$ is some given upper specification limit.

The methodology of any VASP depends on the assumed underlying distribution for the measured variables $X_1, \ldots, X_n$. Here we assume that we deal with a random sample from a normal population with mean $\mu$ and standard deviation $\sigma$. The following discussion will be in terms of a lower specification limit $L$. The corresponding procedure for an upper specification limit $U$ will only be summarized without derivation.

If $L$ is a lower specification limit, then

$$p = p(\mu, \sigma, L) = P_{\mu,\,\sigma}\,(X < L) = P_{\mu,\,\sigma}\left(\frac{X-\mu}{\sigma} < \frac{L-\mu}{\sigma}\right) = \Phi\left(\frac{L-\mu}{\sigma}\right)$$

represents the probability that a given individual item in the population will be defective. Here $\Phi(x)$ denotes the standard normal distribution function. $p$ can be interpreted as the proportion of defective items in the population. It is in the consumer's interest to keep the probability $p$ or proportion $p$ of defective items in the population below a tolerable value $p_1$. Keeping the proportion $p$ low is typically costly for the producer. Hence the producer will try too keep $p$ only so low as to remain cost effective but sufficiently low as not to trigger too many costly rejections. Hence the producer will aim for keeping $p \leq p_0$, where $p_0$ typically is somewhat smaller than $p_1$, in order to provide a sufficient margin between producer and consumer interest.

The consumer's demand that $p \leq p_1$ does not specify how that has to be accomplished in terms of $\mu$ and $\sigma$. The producer can control $p \leq p_0$ by either increasing $\mu$ sufficiently or by reducing $\sigma$, provided $\mu > L$. Reducing $\sigma$ is usually more difficult since various sources of variation have to be controlled more tightly. Increasing $\mu$ is mainly a matter of biasing the process in some way and is usually easier to accomplish. Figure 9 illustrates the two options in the two bottom plots in relation to the population acceptable to the consumer represented in the plot at the top.

For normal data the standard VASP consists in computing $\bar{X}$ and $S$ from the obtained sample of $n$ items and in comparing $\bar{X} - kS$ with $L$ for an appropriately chosen constant $k$. If $\bar{X} - kS \geq L$, the consumer accepts the population from which the sample was drawn and otherwise it is rejected.

Note that rejection or acceptance is not based on the sample proportion of items with $X_i < L$. Such classification would ignore how far above or below $L$ each measurement $X_i$ is. Basing decisions on just such attributes $X_i < L$ or $X_i \geq L$ is much less effective than using the values $X_i$ in their entirety to estimate the underlying normal population and from that get a better idea about $p$ for much smaller sample size. Attribute data should only be used when the direct measurements are not available or not feasible. In that case one needs to employ attribute sampling plans based on the binomial distribution, requiring typically much higher sample sizes.

Before discussing the choice of $k$ in the acceptance criterion $\bar{X} - kS \geq L$ it is appropriate to define the two notions of risk for such a VASP. Due to the random nature of the sample there is some chance that the sample misrepresents the population at least to some extent and thus may induce us to take incorrect action. The consumer's risk is the probability of accepting the population when in fact the proportion $p$ of defectives in the population is greater than the acceptable limit $p_1$. The producer's risk is the probability of rejecting the population when in fact the proportion $p$ of defectives in the population is $\leq p_0$.

It turns out that the probability of acceptance for a given VASP (with its choice of $k$) depends on $\mu, \sigma, L$ only through $p = \Phi((L-\mu)/\sigma)$, the proportion of defectives in the population. This function will thus be denoted by $\gamma(p)$. It is also known as operating characteristic or $OC$–curve of the VASP. $\gamma(p)$ can be expressed in terms of $G_{n-1,\,\delta}(t)$ as follows:

$$\gamma(p) = P_{\mu,\,\sigma}\left(\bar{X} - kS \geq L\right) = P_{\mu,\,\sigma}\left(\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} + \frac{\sqrt{n}(\mu - L)}{\sigma} \geq k\sqrt{n}\frac{S}{\sigma}\right)$$

$$= P_{\mu,\,\sigma}\left(\frac{Z+\delta}{\sqrt{V/(n-1)}} \geq k\sqrt{n}\right) = P(T_{n-1,\,\delta} \geq k\sqrt{n})$$

where the noncentrality parameter

$$\delta = \delta(p) = \frac{\sqrt{n}\,(\mu - L)}{\sigma} = -\sqrt{n}\,\frac{L-\mu}{\sigma} = -\sqrt{n}\,\Phi^{-1}(p) = -\sqrt{n}\,z_p$$

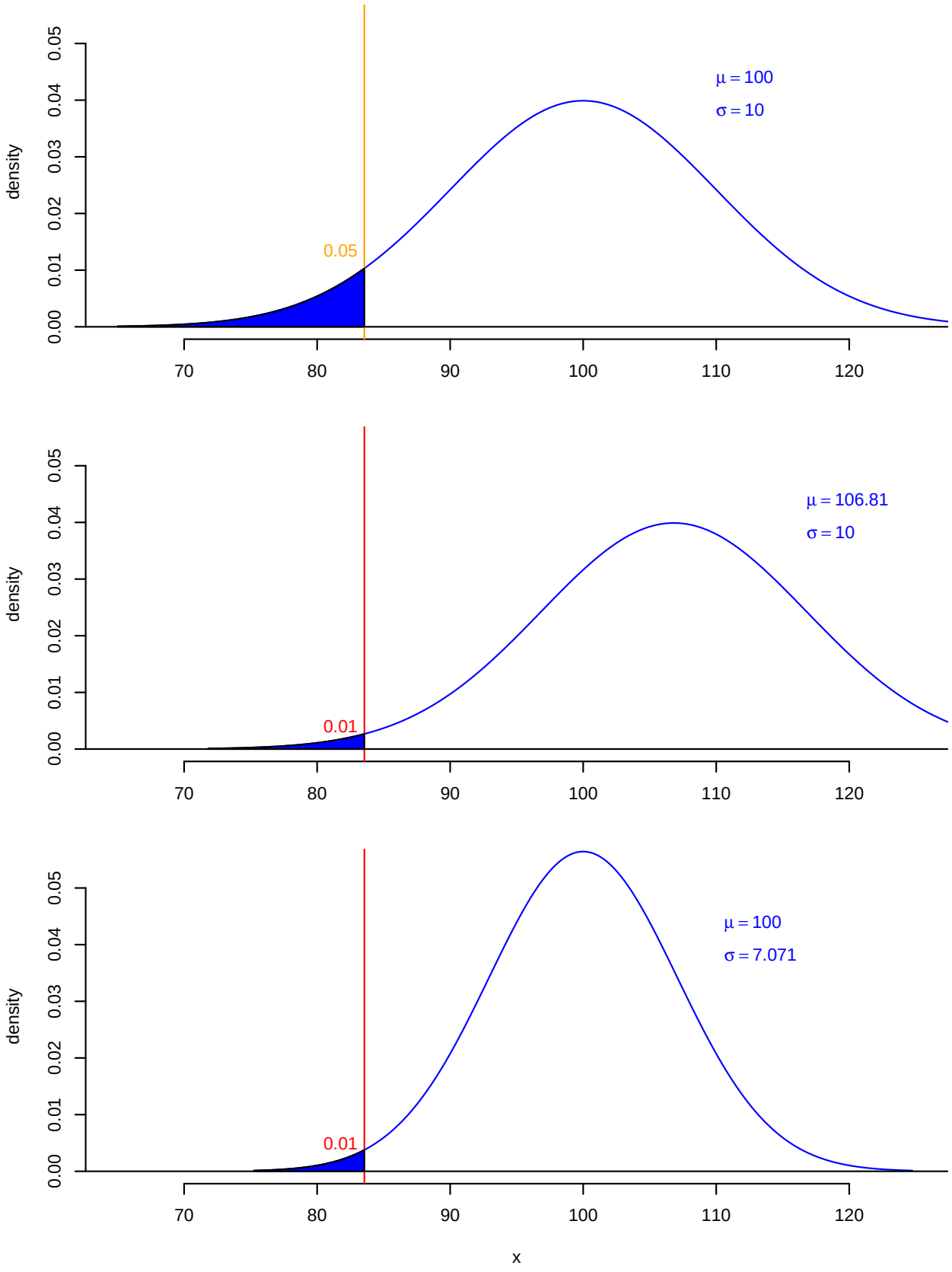



Figure 9: Sampled Population

is a decreasing function of $p$. Hence

$$\gamma(p) = 1 - G_{n-1,\,\delta(p)}(k\sqrt{n})$$

is decreasing in $p$.

The consumer's risk consists of the chance of accepting the population when in fact $p \geq p_1$. In order to control the consumer's risk $\gamma(p)$ has to be kept at some sufficiently small level $\beta$ for $p \geq p_1$. Since $\gamma(p)$ is decreasing in $p$ we need only insure $\gamma(p_1) = \beta$ by proper choice of $k$. The factor $k$ is then found by solving the equation

$$\beta = 1 - G_{n-1,\,\delta(p_1)}(k\sqrt{n}) \qquad \text{for } k, \text{ i.e.,} \qquad k = G^{-1}_{n-1,\,\delta(p_1)}(1-\beta)/\sqrt{n}\ . \tag{4}$$

This is accomplished in R by the command

```
k=qnct(1-beta,n-1,-sqrt(n)*qnorm(p1))/sqrt(n) ,
```

where `beta` $= \beta$ and `p1` $= p_1$. It is customary but not necessarily compelling to choose $\beta = .10$.

Figure 10, produced by the R function `OC.curve.n1`, shows the resulting $k$ and the OC-curve when the sample size is $n = 20$. This solves the problem as far as the consumer is concerned. It does not address the producer's risk requirements. The producer's risk consists of the chance of rejecting the population when in fact $p \leq p_0$. In Figure 10 that risk of rejecting the population is seen to be as high as .3575 when $p_0 \leq .01$.

The probability of rejecting the population is $1 - \gamma(p)$, which is maximal over $p \leq p_0$ at $p_0$. Hence the producer would want to limit this maximal risk $1 - \gamma(p_0)$ by some value $\alpha$, customarily chosen to be .05. Note that $\alpha$ and $\beta$ must satisfy the constraint $\alpha + \beta < 1$. Thus the producer is interested in ensuring that

$$\alpha = 1 - \gamma(p_0) = G_{n-1,\,\delta(p_0)}(k\sqrt{n}) \tag{5}$$

Solving this for $k$, i.e., using `beta` $= \beta$ and `p0` $= p_0$

$$\texttt{k} = \texttt{qnct}(\texttt{alpha}, \texttt{n}-1, -\texttt{sqrt}(\texttt{n}) * \texttt{qnorm}(\texttt{p0}))/\texttt{sqrt}(\texttt{n})\ ,$$

will typically lead to a different choice from that obtained in (4) leaving us with a conflict. This is illustrated in Figure 11, produced by the R function `OC.curve.n0`, which shows the resulting $k$ and the OC-curve when the sample size is again $n = 20$.

Note that in Figure 10 we accept when $\bar{X} - 2.208 \times S \geq L$ (keeping the consumer risk at $\beta = .10$) while in Figure 11 we accept more readily when $\bar{X} - 1.749 \times S \geq L$ and thus keeping the producer risk at $\alpha = .05$.

This conflict of having two different values of $k$, depending on whose interest is being served, can be resolved by leaving the sample size $n$ flexible so that there are two control parameters, $n$ and $k$, which can be used to satisfy the two conflicting goals. One slight problem is that $n$ is an integer and so it may not be possible to satisfy both equations (4) and (5) exactly.

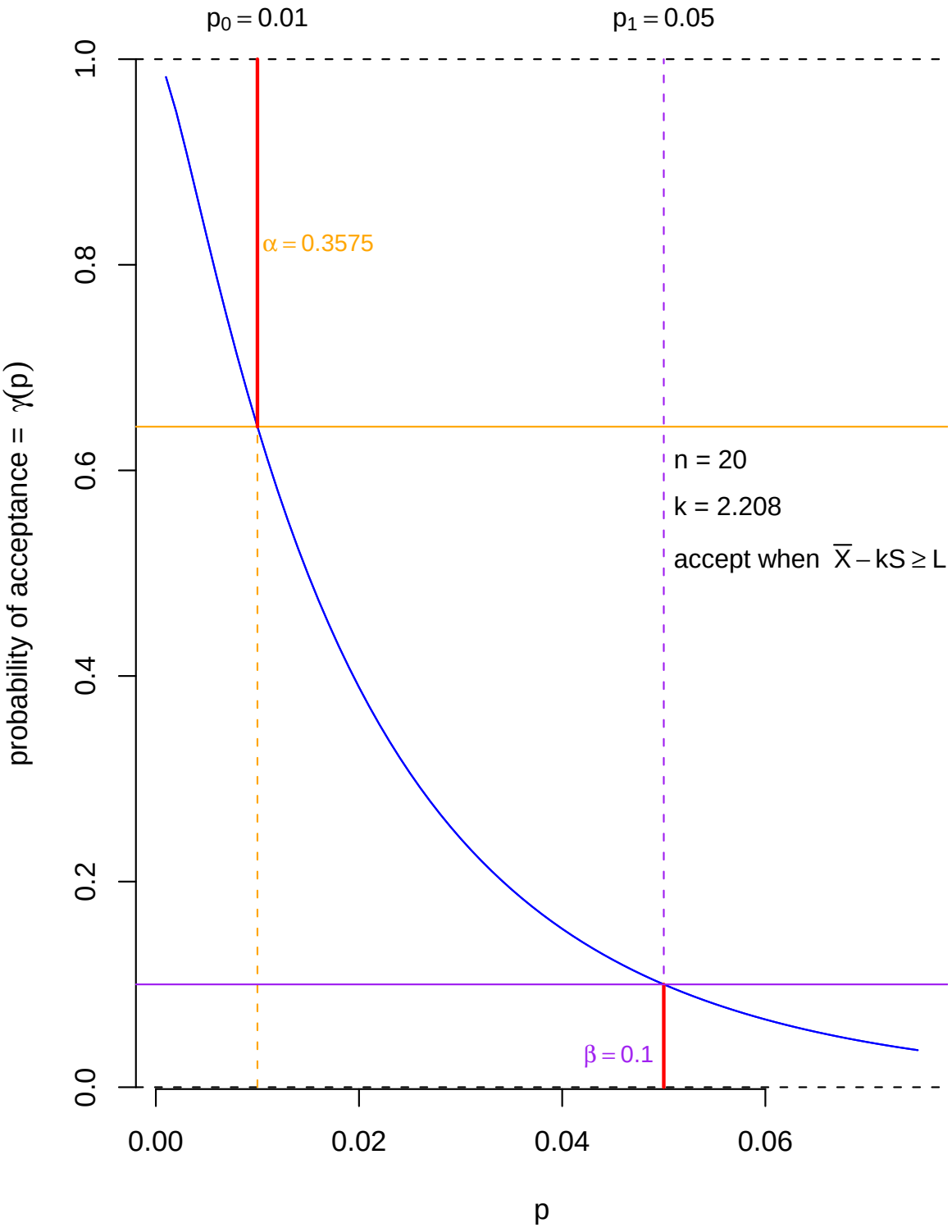



Figure 10: Operating Characteristic Curve
Protecting the Consumer Risk

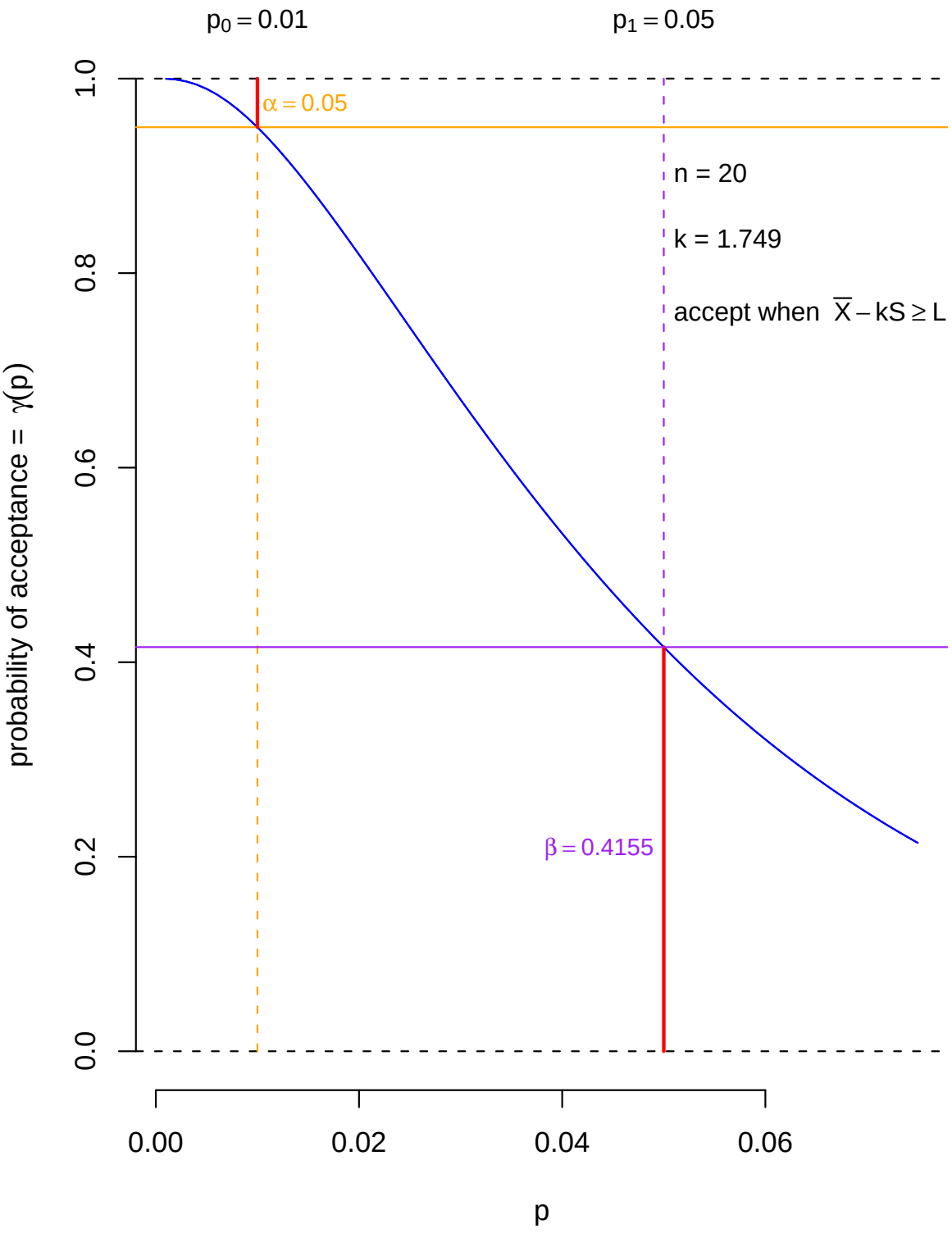



Figure 11: Operating Characteristic Curve
Protecting the Producer Risk

What one can do instead is the following: For a given value $n$ find $k = k(n)$ to solve (4). If that $k(n)$ also yields

$$\alpha \geq G_{n-1,\,\delta(p_0)}(k(n)\sqrt{n}) \;, \tag{6}$$

then this sample size $n$ was possibly chosen too high and a lower value of $n$ should be tried. If we have

$$\alpha < G_{n-1,\,\delta(p_0)}(k(n)\sqrt{n}) \;,$$

then $n$ was definitely chosen too small and a larger value of $n$ should be tried next. Through iteration one can arrive at the smallest sample size $n$ such that $k(n)$ and $n$ satisfy both (4) and (6). This iteration process will lead to a solution provided $p_0 < p_1$. If $p_0$ and $p_1$ are too close to each other, very large sample sizes will be required. Note that the search for the minimal sample size $n$ does not involve $L, \mu$ and $\sigma$. Only $p_0, p_1, \alpha$ and $\beta$ are required. Such a process is carried out by the R function `OC.curve` which also produces the plot in Figure 12, indicating the appropriate choice for $n$ and $k$.

In the case of an upper specification limit $U$ we accept the lot or population whenever

$$\bar{X} + kS \leq U \;.$$

By rewriting $X > U$ as $X' = -X < -U = L$ this reduces to the previous case. Then $S' = S$ and

$$\bar{X} + kS \leq U \quad \Longleftrightarrow \quad -\bar{X} - kS \geq -U \quad \Longleftrightarrow \quad \bar{X}' - kS' \geq L$$

and $p = P(X > U) = P(X' < L)$. The same $k$ and $n$ as before suffice as solution as long as we identify $p = P(X > U)$ with $p = P(X' < L)$, i.e., specify only this risk $p$ of a population item being defective. Recall that the values of $L, \mu$ and $\sigma$ did not enter explicitly in our derivation.

We point out that the VASP does not say how the producer accomplishes the value $p \leq p_0$. This is usually based on extensive testing or the producer's broad experience. This may lead to calculating upper confidence bounds for $P(X < L)$ based on sufficient data. This is addressed in a later section.

Also, the consumer cannot set $p_1$ arbitrarily low since there may not be a producer that will deliver that quality or will deliver it only at exorbitant costs.

We compare the previously discussed Variables Acceptance Sampling Plan (VASP) with the corresponding Attributes Acceptance Sampling Plan (AASP) in order to understand the effect on the required sample size $n$ when all requirements are kept at the same levels. Figure 13, produced by the R function `OC.binom`, shows the OC-curve of the AASP.

In an AASP the number $X$ of defective items is counted and the population is accepted when $X \leq k$, where $k$ and the smallest sample size $n$ are determined such that for given $p_0 < p_1$ and $\alpha > 0$, $\beta > 0$ with $\alpha + \beta < 1$ we have

$$P_{p_1}(X \leq k) \leq \beta \qquad \text{and} \qquad P_{p_0}(X \leq k) \geq 1 - \alpha \;. \tag{7}$$

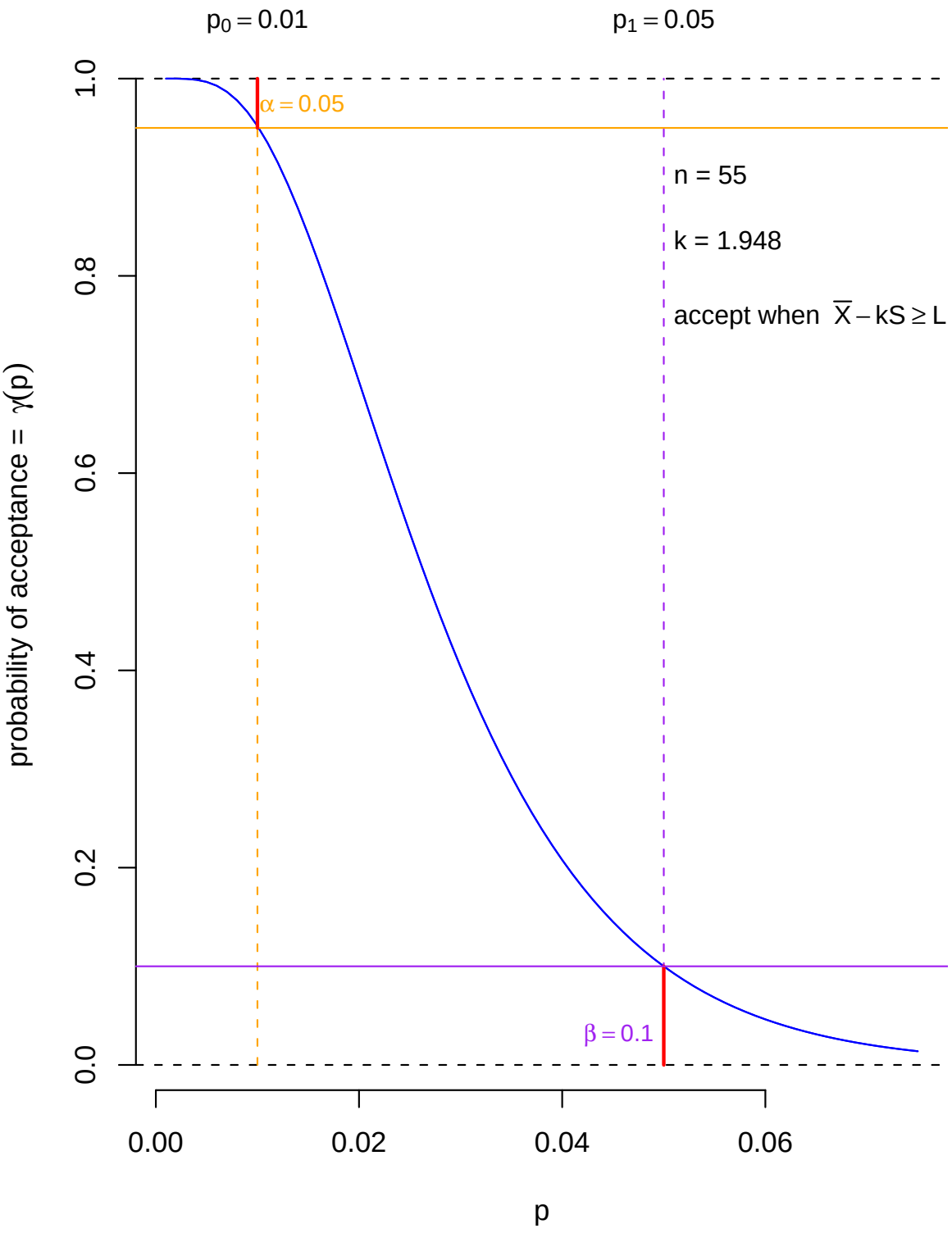



Figure 12: Operating Characteristic Curve
Protecting Both Risks

For Figure 13 we chose the same requirements $p_1 = .05$, $p_0 = .01$, $\beta = .1$ and $\alpha = .05$ as were used in Figure 12 to enable a clear comparison of the involved sample sizes. For the VASP we had $n = 55$ while for the attributes sampling plan we have $n = 132$, which is considerably higher.

In finding $n$ we start out with an $n$ suggested by the normal approximation to $X$ (with continuity correction)

$$P_p\left(X \le k\right) \approx \Phi\left(\frac{k + .5 - np}{\sqrt{np(1-p)}}\right) = \gamma(p) \; .$$

The requirements

$$\gamma(p_0) = 1 - \alpha \qquad \text{and} \qquad \gamma(p_1) = \beta$$

translate to

$$\frac{k + .5 - np_0}{\sqrt{np_0(1-p_0)}} = z_{1-\alpha} = -z_\alpha \qquad \text{and} \qquad \frac{k + .5 - np_1}{\sqrt{np_1(1-p_1)}} = z_\beta$$

or

$$k + .5 - np_0 = -z_\alpha\sqrt{np_0(1-p_0)} \qquad \text{and} \qquad k + .5 - np_1 = z_\beta\sqrt{np_1(1-p_1)}$$

and by subtracting the first equation from the second we cancel out $k$ and get a single equation involving $n$

$$np_0 - np_1 = z_\beta\sqrt{np_1(1-p_1)} + z_\alpha\sqrt{np_0(1-p_0)}$$

or

$$\sqrt{n}\,(p_0 - p_1) = z_\beta\sqrt{p_1(1-p_1)} + z_\alpha\sqrt{p_0(1-p_0)}$$

which yields

$$n = \left(\frac{z_\beta\sqrt{p_1(1-p_1)} + z_\alpha\sqrt{p_0(1-p_0)}}{p_1 - p_0}\right)^2 \quad \text{rounded up to the next integer.}$$

For this $n$ we find `k=qbinom(beta,n,p1)` which is the smallest $k$ such that $P_{p_1}(X \le k) \ge \beta$, where `beta`=$\beta$ and `p1`=$p_1$. If this $k$ gives $P_{p_1}(X \le k) > \beta$ we reduce $k$ by 1 and leave it alone otherwise. For given $n$ this $k$ is then the largest value for which $P_{p_1}(X \le k) \le \beta$. We then evaluate $P_{p_0}(X \le k)$. If it is larger than $1 - \alpha$ we have $n$ possibly too large and we reduce $n$ by 1. If $P_{p_0}(X \le k)$ is smaller than $1 - \alpha$ we have $n$ too small and we increase $n$ by 1. Recalculating $k$ for each new $n$ and rechecking $P_{p_0}(X \le k)$ against $1 - \alpha$ we iteratively find the smallest $n$ for which we satisfy both constraints (7). Although the original algorithm implemented the above starting value in the R function `OC.binom` it was later modified to start at $n = 1$, which is cruder but gets the right answer for sure. The reason for this is that for an OC-curve corresponding to a largest $k$ choice with OC($p_1 \le \beta$ the achieved OC($p_0$) is not monotone in $n$, due to the discrete nature of $X$. OC($p_0$) shows a pronounced zigzag behavior with respect to $n$, see Figure 14 as an illustration. This complicates the straightforward search for the smallest $n$ satisfying the above requirements.

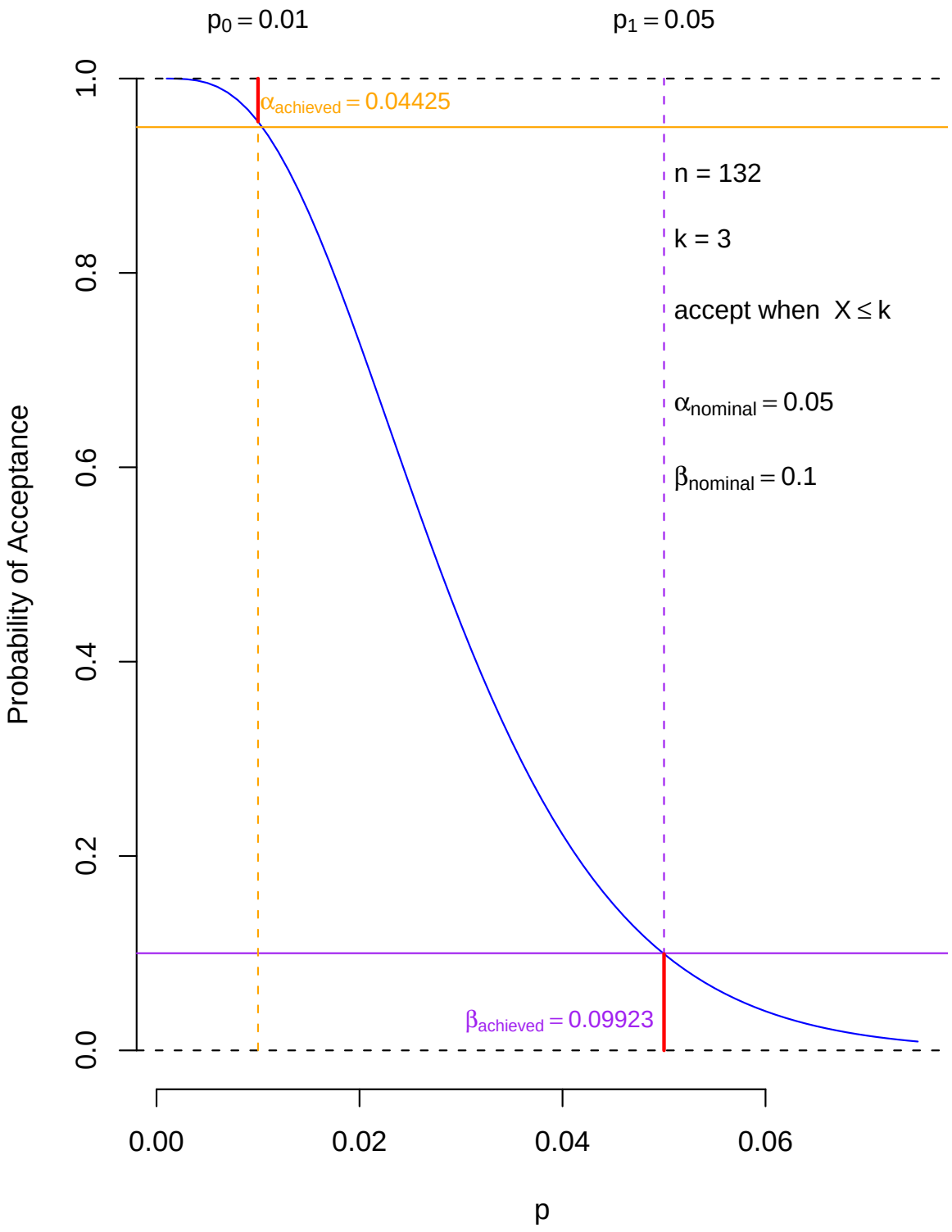



Figure 13: Operating Characteristic Curve
for Attribute Acceptance Sampling Plan

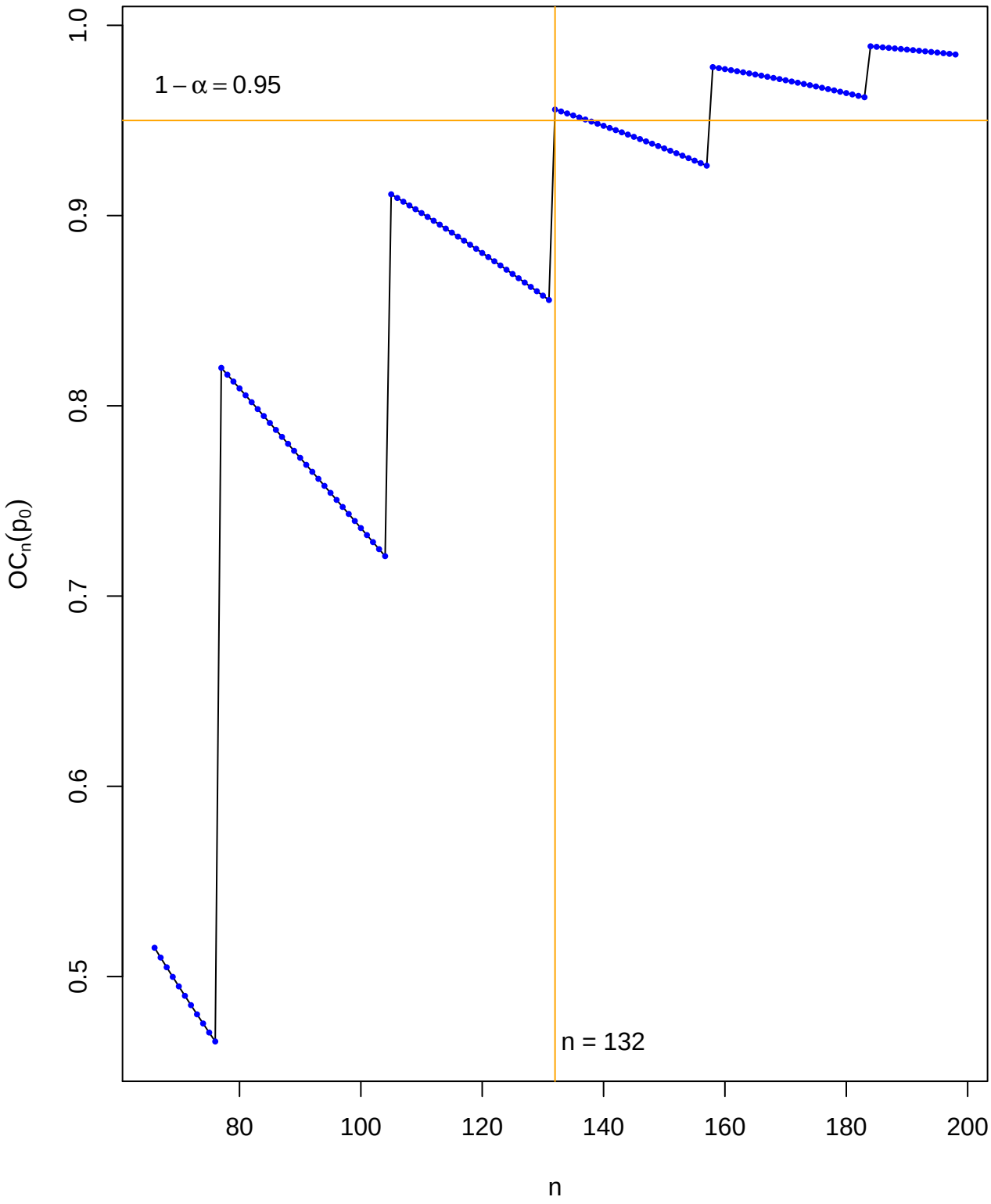



Figure 14: $\mathrm{OC}_n(p_0)$ for AASP with $\mathrm{OC}_n(p_1) \leq \beta = .10$ Tests

## 6 Tolerance Bounds or Tolerance Limits

Tolerance bounds or tolerance limits are lower or upper confidence bounds on percentiles or quantiles of a population, here again assumed to be normal. The discussion will mainly focus on lower confidence bounds. The upper bounds fall out immediately from the lower bounds by a simple switch to the complementary confidence level as explained below.

The $p$-percentile or $p$-quantile $x_p$ of a normal population with mean $\mu$ and standard deviation $\sigma$, denoted by $\mathcal{N}(\mu, \sigma^2)$, can be expressed as

$$x_p = \mu + z_p\, \sigma \ ,$$

where $z_p = \Phi^{-1}(p)$ is the $p$-quantile of the standard normal distribution.

Using a sample $X_1, \ \ldots, \ X_n$ taken from this population we estimate $\mu$ and $\sigma$ again by using $\bar{X}$ and $S$. The lower confidence bound for $x_p$ is then computed as $\hat{x}_{p,L}(\gamma) = \bar{X} - kS$ where $k$ is determined to achieve the desired confidence level $\gamma$, namely so that for all $(\mu, \ \sigma)$ we have

$$P_{\mu,\,\sigma}(\bar{X} - kS \le x_p) = \gamma \ ,$$

which has the same form as equation (1).

By complementation this yields immediately that for all $(\mu, \ \sigma)$

$$P_{\mu,\,\sigma}(\bar{X} - kS \ge x_p) = 1 - \gamma \ ,$$

i.e., $\bar{X} - kS$ also serves as an upper bound for $x_p$ with confidence level $1 - \gamma$. Of course, to get a confidence level of .95 for such an upper bound one would choose $\gamma = .05$ in the above interpretation of $\bar{X} - kS$ as upper bound.

Invoking equation (2) we have

$$P_{\mu,\,\sigma}(\bar{X} - kS \le x_p) = P(T_{n-1,\,\delta} \le \sqrt{n}k) = G_{n-1,\,\delta}(\sqrt{n}k) \ ,$$

where $\delta = -\sqrt{n}(x_p - \mu)/\sigma = -\sqrt{n}z_p$. Hence $k$ is determined by solving the following equation for $k$:

$$G_{n-1,\,\delta}(\sqrt{n}k) = \gamma \ .$$

In R this is done by invoking the command

$$\texttt{k} = \texttt{qnct}(\texttt{gam}, \texttt{n} - 1, -\texttt{sqrt}(\texttt{n}) * \texttt{qnorm}(\texttt{p}))/\texttt{sqrt}(\texttt{n}) \ ,$$

where gam=$\gamma$. Avoid the variable name gamma in R since it is the intrinsic $\Gamma$-function.

In structural engineering the 95% lower bounds for $x_{.01}$ and $x_{.10}$ are called $A$- and $B$-Allowables, respectively, and are mainly used to limit material strength properties from below. In the lumber industry the interest is in 75% lower bounds for $x_{.05}$, see page 4 of

https://www.aitc-glulam.org/shopcart/Pdf/aitc_402-2005.pdf

402.4.8. Beam Performance. The beam strength 5% tolerance limit
with 75% confidence determined in accordance with ASTM D2915
shall be a minimum of 2.1 times the design value for the beam.
....

As an illustration we will use some data from MIL-HDBK-5J[1], see
http://www.weibull.com/mil_std/mil_hdbk_5j.pdf.
In particular, we will use the TUS (tensile ultimate strength) data set, designated as Group 5 on page 9-165. It consists of $n = 100$ values, measured in KSI (1000 pounds per square inch) and is contained in the referenced R work space as m5dat5.
The normal QQ-plot of this data set is shown in Figure 15 (produced by m5dat5.qqnorm) and exhibits no significant deviation from normality. Formal tests for normality, Lilliefors (Kolmogorov-Smirnov), Cramér-von Mises, and Anderson-Darling, confirm this with $p$-values above .63 for all three discrepancy metrics. These tests are available as part of the nortest package. Download nortest_1.0.zip from the class web site to the directory that houses your R work space. Under the R Packages menu item install this package. This installation needs to be done only once on your computer for the installed version of R. After this installation you need to invoke library(nortest) in any R session that wants to use the functions in the package nortest. These functions are lillie.test, cvm.test and ad.test and you get documentation on them by placing a ? in front of the respective function names, e.g., ?lillie.test.
The sample mean and standard deviation are $\bar{X} = 145$ and $S = 4.469965$, respectively. The $k$-factors for $A$- and $B$-allowables are respectively

$$k_A = \texttt{qnct}(.95, 99, -\texttt{sqrt}(100) * \texttt{qnorm}(.01))/\texttt{sqrt}(100) = 2.683957$$

and

$$k_B = \texttt{qnct}(.95, 99, -\texttt{sqrt}(100) * \texttt{qnorm}(.1))/\texttt{sqrt}(100) = 1.526749$$

so that the $A$-and $B$-allowables are

$$A = \hat{x}_{.01,L}(.95) = \bar{X} - k_A \times S = 145 - 2.683957 \times 4.469965 = 133.0028$$

and

$$B = \hat{x}_{.10,L}(.95) = \bar{X} - k_B \times S = 145 - 1.526749 \times 4.469965 = 138.1755 \ .$$

Figure 16 shows these allowables in relation to the data and their histogram.

[1]Note that this file is about 68.5MB and consists of 1733 pages

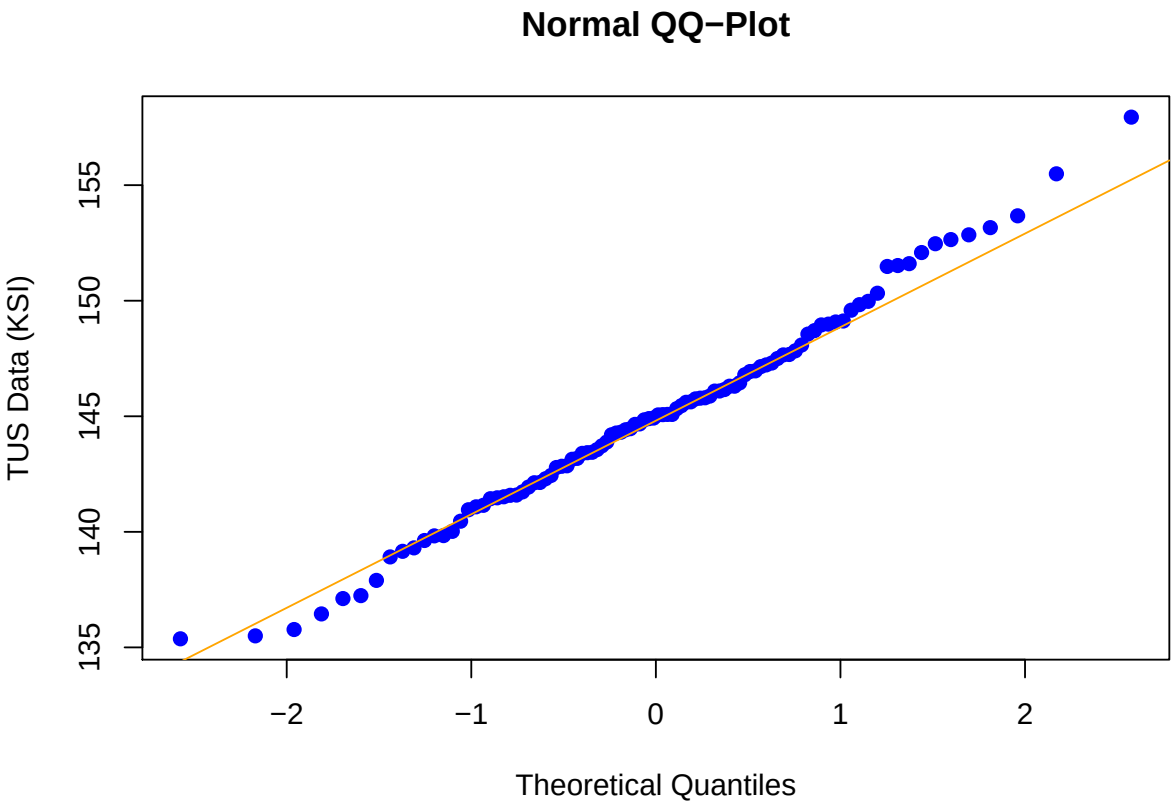



Figure 15: Normal QQ-Plot of MIL-HDBK-5J Group 5 Data

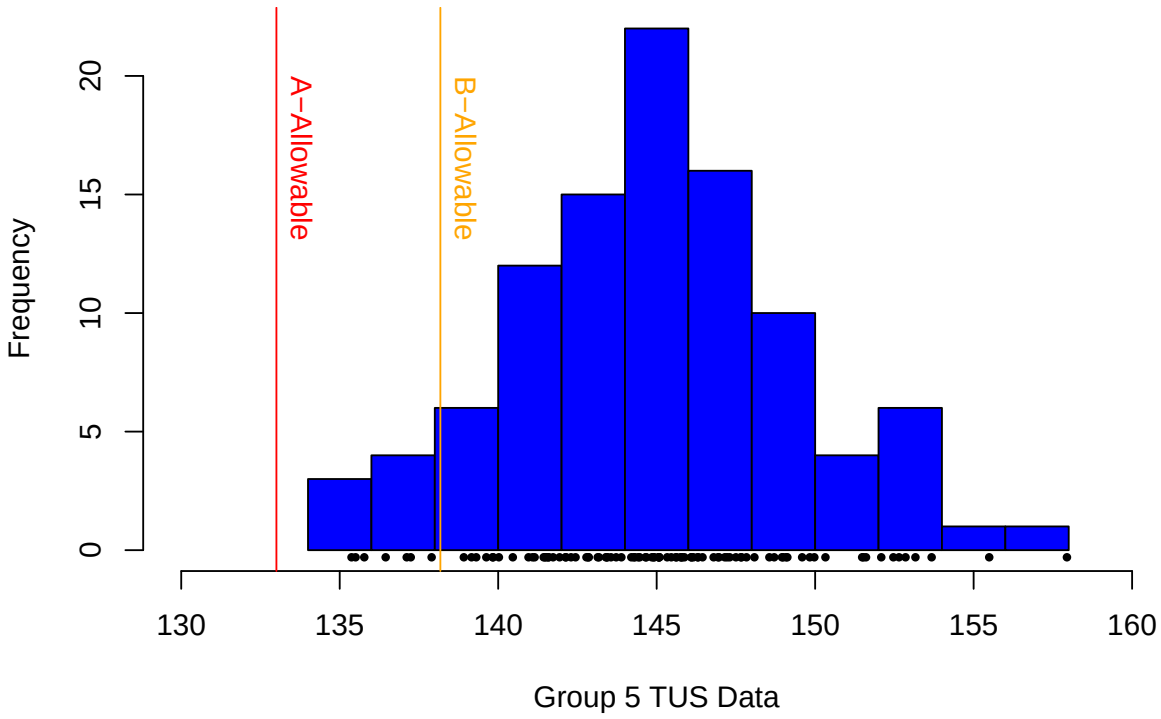



Figure 16: MIL-HDBK-5J Group 5 Data with $A$- and $B$-Allowables

## 7 Tail Probability Confidence Bounds

Of interest here are the tail probabilities of a normal population with mean $\mu$ and standard deviation $\sigma$. For a given threshold value $x_0$ one is interested in the tail probability

$$p = p(x_0) = p(x_0, \mu, \sigma) = P_{\mu,\,\sigma}(X \leq x_0) = \Phi\left(\frac{x_0 - \mu}{\sigma}\right) .$$

In the context of variables acceptance sampling plans this came up as the probability $p = P(X < L)$ of an item being defective. Upper bounds for such probabilities $p$ could give a producer the needed assurance of having a proportion of defectives $\leq p_0$, the value used in setting up the VASP.

Although $\hat{p} = \Phi((x_0 - \bar{X})/S)$ is a natural but biased estimate of $p$ the construction of confidence bounds is not so obvious. It seems that there should be a link between quantile bounds and tail probability bounds, and there is. However, in constructing such tail probability bounds we will take a more direct approach and revisit the link later.

First we will discuss the generic relationships between upper and lower bounds for left and right tail probabilities $p$ and $q = 1 - p$.

If $\hat{p}_U(\gamma)$ denotes an upper bound for $p$ with confidence level $\gamma$, i.e., for all $(\mu,\ \sigma)$

$$P_{\mu,\,\sigma}(\hat{p}_U(\gamma) \geq p) = \gamma ,$$

then we also have for all $(\mu,\ \sigma)$

$$P_{\mu,\,\sigma}(\hat{p}_U(\gamma) \leq p) = 1 - \gamma ,$$

so that $\hat{p}_U(\gamma)$ can also serve as a lower bound $\hat{p}_L(1-\gamma) = \hat{p}_U(\gamma)$ for $p$ with confidence level $1 - \gamma$.

If the upper tail probability $q = 1 - p$ of the normal distribution is of interest, then $1 - \hat{p}_U(\gamma)$ will serve as lower bound for $q$ with confidence level $\gamma$ and thus as an upper bound for $q$ with confidence level $1 - \gamma$. Thus it suffices to limit any further discussion to upper confidence bounds for $p$.

In deriving these upper bounds we will use the following result known as the *Probability Integral Transformation*. We state and prove it here in simplified form, using the convenient additional but unnecessary assumption of strict monotonicity, see Lehmann & Romano (2005), p. 97.

**Lemma:** If $X$ is a random variable with continuous and strictly increasing distribution function $F(t) = P(X \leq t)$, then the random variable $U = F(X)$ is uniformly distributed over $[0, 1]$, i.e., $P(U \leq u) = u$ for $0 \leq u \leq 1$.

**Proof:**

$$P(U \leq u) = P(F(X) \leq u) = P(X \leq F^{-1}(u)) = F(F^{-1}(u)) = u \qquad \text{q.e.d.}$$

As a start for constructing upper bounds for $p$ consider

$$\frac{\sqrt{n}(x_0 - \bar{X})}{S} = \frac{\sqrt{n}(x_0 - \mu)/\sigma + \sqrt{n}(\mu - \bar{X})/\sigma}{S/\sigma} = T_{n-1,\,\delta} \ ,$$

and note that $Z' = \sqrt{n}(\mu - \bar{X})/\sigma$ and $Z = \sqrt{n}(\bar{X} - \mu)/\sigma = -Z'$ have the same standard normal distribution. Here $\delta = \sqrt{n}(x_0 - \mu)/\sigma = \sqrt{n}\Phi^{-1}(p)$ is an increasing function of $p$. By the above Lemma the random variable

$$U = G_{n-1,\,\delta}\left(\frac{\sqrt{n}(x_0 - \bar{X})}{S}\right) = G_{n-1,\,\delta}\left(T_{n-1,\,\delta}\right)$$

has a uniform distribution over the interval $(0, 1)$. Such a function $U$ of the sample data and the unknown parameters is called a pivot when its distribution is completely known, as it is the case here. The concept of pivots is often employed in constructing confidence sets.
We have

$$\gamma = P(U \geq 1 - \gamma) \ . \tag{8}$$

and since $G_{n-1,\,\delta}(t)$ is decreasing in $\delta$ we have

$$U = G_{n-1,\,\delta}\left(\sqrt{n}(x_0 - \bar{X})/S\right) \geq 1 - \gamma \qquad \Longleftrightarrow \qquad \delta \leq \hat{\delta} \ ,$$

where $\hat{\delta}$ solves

$$G_{n-1,\,\hat{\delta}}\left(\sqrt{n}(x_0 - \bar{X})/S\right) = 1 - \gamma \ . \tag{9}$$

Hence $\hat{\delta}$ is an upper confidence bound for $\delta = \sqrt{n}\Phi^{-1}(p)$ with confidence level $\gamma$. Since

$$\hat{\delta} \geq \delta = \sqrt{n}\Phi^{-1}(p) \qquad \Longleftrightarrow \qquad \hat{p}_U = \hat{p}_U(\gamma) = \hat{p}_U(x_0, \gamma) \stackrel{\text{def}}{=} \Phi(\hat{\delta}/\sqrt{n}) \geq \Phi(\delta/\sqrt{n}) = p \ ,$$

$\hat{p}_U$ is the desired upper confidence bound for $p$ with confidence level $\gamma$.
This upper confidence bound is found by invoking the following R command

$$\hat{p}_U(x_0, \gamma) = \texttt{pnorm}(\texttt{del.nct}(\texttt{sqrt}(\texttt{n}) * (\texttt{x0} - \texttt{Xbar})/\texttt{S}, 1 - \texttt{gam}, \texttt{n} - 1)/\texttt{sqrt}(\texttt{n}))$$

where `gam`$= \gamma$, `Xbar`$= \bar{X}$, `S`$= S$, and `x0`$= x_0$. Again avoid `gamma` as a variable name.
We point out that the coverage probability statement in (8) holds for any $(\mu, \sigma)$, which enter through $U$ in two-fold form, namely through $\delta$ in $G_{n-1,\,\delta}$ and through the joint distribution of $\bar{X}$ and $S$ in $\sqrt{n}(x_0 - \bar{X})/S$. This means that the coverage probability is constant in $\mu$ and $\sigma$ and thus equals the confidence coefficient or the minimum coverage probability $\bar{\gamma}$. The same comment applies to tolerance bounds. Compare this behavior with the more complex behavior seen earlier in the context of confidence bounds for discrete distributions, such as the Poisson, Binomial and Hypergeometric distributions.

It turns out that the upper bounds for left tail probabilities $p(x) = P(X \leq x)$ are just the inverse to the lower bounds for the $x_{p(x)}$-quantile and vice versa. This is illustrated in the normal probability plot of Figure 17 which was produced by the function `normal.paper`. Note that this function illustrates the use of the function `set.seed` which makes sure that the same seed is used in each simulation. The seed `iseed` is an argument to `normal.paper`, with default `iseed=25`.

Using a random sample of size $n = 30$ from $\mathcal{N}(\mu, \sigma^2)$ with $\mu = 100$ and $\sigma = 10$, Figure 17 shows a QQ-plot or normal probability plot of the sample. In such a plot the $i^{\text{th}}$ smallest sample value $X_{(i)}$ is plotted against the standard normal $p_i$-quantile $z_{p_i}$, with $p_i = (i - .5)/n$, $i = 1, 2, \ldots, n$, i.e., we plot the sample quantiles against the corresponding standard normal quantiles from which derives the term "QQ-plot." However, here the markings on the abscissa are given in terms of $p$, which explains the parallel term "normal probability plot." One expects that the $i^{\text{th}}$ sample quantile $X_{(i)}$ is a reasonable estimate of the corresponding population quantile $x_{p_i} = \mu + \sigma z_{p_i}$, i.e., $X_{(i)} \approx x_{p_i}$, and since the latter is a linear function of $z_{p_i}$ one would expect to see a roughly linear pattern when plotting $X_{(i)}$ vs $z_{p_i}$. This is the basic idea behind the normal QQ-plot and its informal diagnostic appeal for judging data normality.

The line through the data is just $\bar{X} + z_p S$. The curve below that line represents either the 95% lower bound for $x_p$ when read sideways from the curve at the vertical $p$ intercept, or it represents the 95% upper bound $\hat{p}_U(x)$ for the left tail probability $p(x)$ when read vertically down from the curve at the horizontal $x$ intercept. It would have been possible to introduce these upper bounds $\hat{p}_U(x)$ using this inverse relationship but we preferred the direct approach given above. It leads directly to the solution path via `del.nct`, rather than inverting the above curve relationship. The function `del.nct` involves a single root solving step. Inverting the curve involves two root solving steps, one for computing the curve via `qnct` and one for inverting it. The proof for the equivalence of the pivot based approach and that using the curve inversion it relegated to Appendix A.

We note that the binomial upper bound for $P(X \leq 80) = \Phi((80-100)/2) = 0.02275$ would be based on the zero count of observations $\leq 80$, i.e., it comes out to be `qbeta(.95,1,30)=0.09503385`. This is lower than $\hat{p}_U(80) = 0.1109$ as obtained from $\bar{X}$ and $S$, presumably because the lowest sample value is somewhat high compared to what is suggested by the line $\bar{X} + z_p S$. If it had been $\leq 80$ we would get an upper bound $\geq$ `qbeta(.95, 2, 29)` $=$ `0.1485961`.

This should serve as an example for two comments. Confidence bounds based on the same data but using different methods are typically different. Furthermore, even if method $A$ (based on $\bar{X}$ and $S$) is generally superior to method $B$ (binomial method), it can happen (as in this instance) that the bound produced by $B$ is "better" than the bound produced by $A$. Both upper bounds are above the true target 0.02275 but the binomial bound happens to be closer.

Finally, we point out that the 95% confidence curve has to be interpreted point-wise, i.e., the probability for several such upper (or lower) bounds simultaneously covering their respective targets is $< .95$.

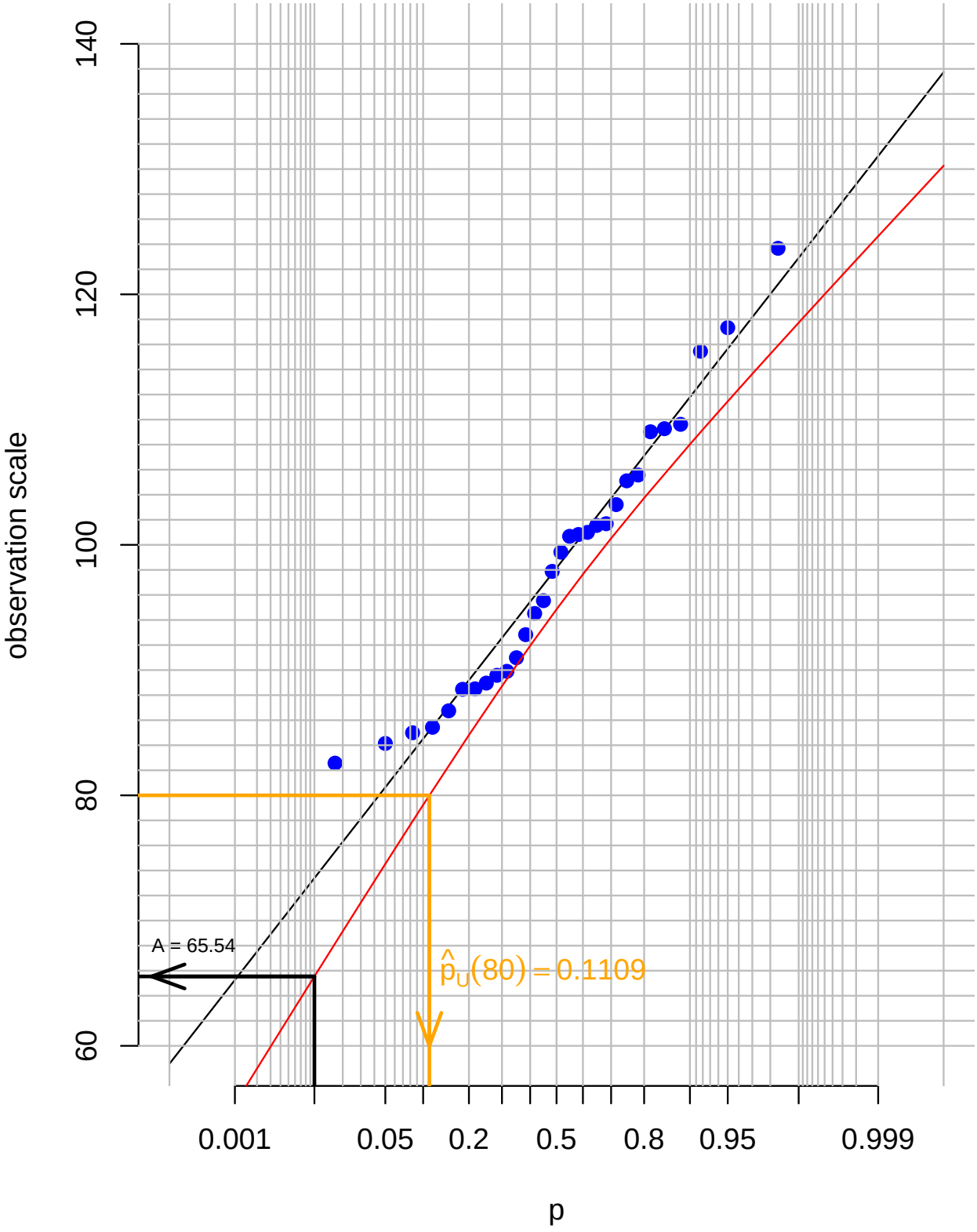



Figure 17: Normal Probability Plot with Confidence Curve for $\hat{x}_L(p)$ and $\hat{p}_U(x)$

# 8 Bounds for Process Control Capability Indices

The process control capability indices $C_L$, $C_U$ and $C_{pk}$ are relatively new in quality control applications. They are defined as

$$C_L = \frac{\mu - x_L}{3\sigma} \ , \quad C_U = \frac{x_U - \mu}{3\sigma} \quad \text{and} \quad C_{pk} = \min(C_L, C_U) \ ,$$

where $x_L$ and $x_U$ are given lower and upper specification limits. Again it is assumed that the process output is governed by a normal distribution with mean $\mu$ and standard deviation $\sigma$. Values $C_L \geq 1$, $C_U \geq 1$ and $C_{pk} \geq 1$ indicate that the process output is at least $3\sigma$ units on the safe side from any specification limit, since

$$C_L \geq 1 \iff \mu{-}3\sigma \geq x_L \ , \ \ C_U \geq 1 \iff \mu{+}3\sigma \leq x_U \ , \ \ C_{pk} \geq 1 \iff x_L \leq \mu{-}3\sigma < \mu{+}3\sigma \leq x_U \ .$$

As Figure 18 shows, there are many combinations of $(\mu, \sigma)$ for which these indices are are 1. As $\sigma$ increases the further the mean has to move away from its respective one-sided specification limit. This does not work when we have a specification interval. In order to have $C_{pk} \geq 1$ we must have $6\sigma \leq x_U - x_L$. This latter aspect is addressed by a different capability index, namely $C_p = (x_U - x_L)/(6\sigma)$. Confidence bounds for it can be obtained from confidence bounds for $\sigma$ based on the $\chi^2$ distribution. Another index, originating in Japan, is $k = 2|(U + L)/2 - \mu|/(U - L)$ and it links $C_{pk}$ and $C_p$ via $C_{pk} = C_p(1 - k)$. For an extensive treatment of capability indices we refer to Kotz and Johnson (1993) and Kotz and Lovelace (1998).

Typically the parameters $\mu$ and $\sigma$ are unknown and only limited sample data, say $X_1, \ldots, X_n$, are available from this population. The next 3 subsections show how to obtain lower confidence bounds for these indices. Lower bounds are of primary interest here since it is typically desired to show that the process capability index meets at least a certain threshold, say 1 or 4/3.

## 8.1 Lower Confidence Bounds for $C_L$

A natural estimate for $C_L$ is $\widehat{C}_L = (\bar{X} - x_L)/3S$ and it is the basis for constructing $100\gamma\%$ lower confidence limits for $C_L$. We have

$$\begin{aligned} P\left(\widehat{C}_L \leq k\right) &= P\left(\frac{\bar{X} - x_L}{3S} \leq k\right) \\ &= P\left(\frac{\sqrt{n}(\bar{X} - \mu)/\sigma + \sqrt{n}(\mu - x_L)/\sigma}{S/\sigma} \leq 3\sqrt{n}k\right) = P\left(T_{n-1,3\sqrt{n}C_L} \leq 3\sqrt{n}k\right) . \end{aligned}$$

We define $k = k(C_L)$ as that unique number which for given $C_L$ solves

$$P\left(T_{n-1,3\sqrt{n}C_L} \leq 3\sqrt{n}k(C_L)\right) = \gamma \ .$$

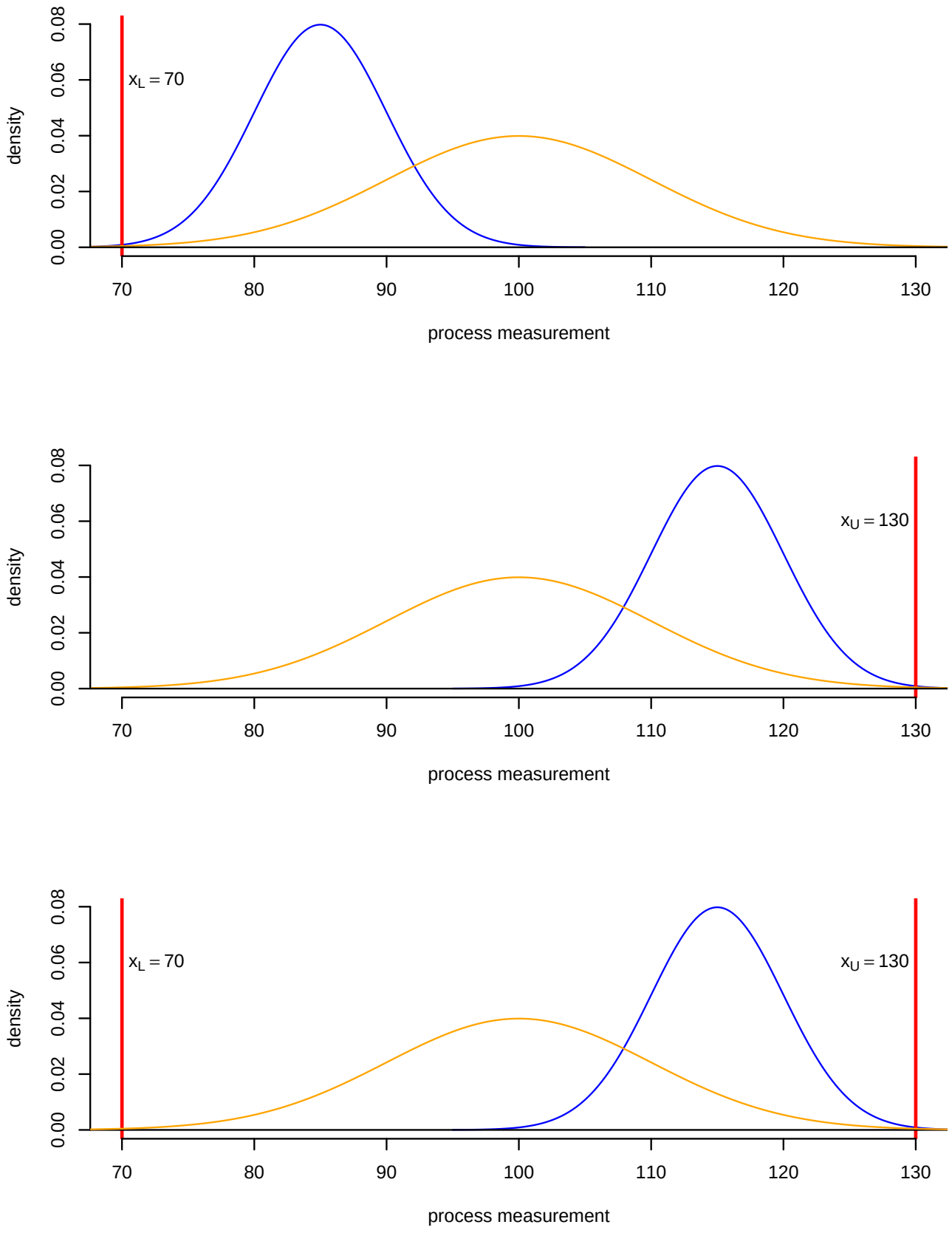



Figure 18: Process Measurements in Relation to Specification Limits

From the previously cited monotonicity properties of the noncentral $t$-distribution we know that $k(C_L)$ is a strictly increasing function of $C_L$. Thus we have

$$\gamma = P\left(\widehat{C}_L \le k(C_L)\right) = P\left(k^{-1}(\widehat{C}_L) \le C_L\right)$$

and we can treat $\widehat{B}_L = k^{-1}\left(\widehat{C}_L\right)$ as a $100\gamma\%$ lower confidence bound for $C_L$.

It remains to show how $\widehat{B}_L$ is actually computed for each such observed value $\widehat{c}_L$ of $\widehat{C}_L$.

In the defining equation for $k(C_L)$ take $C_L = k^{-1}(\widehat{c}_L)$ and rewrite that defining equation as follows:

$$\gamma = P\left(T_{n-1,3\sqrt{n}k^{-1}(\widehat{c}_L)} \le 3\sqrt{n}k\left(k^{-1}(\widehat{c}_L)\right)\right) = P\left(T_{n-1,3\sqrt{n}k^{-1}(\widehat{c}_L)} \le 3\sqrt{n}\widehat{c}_L\right) = \gamma\ .$$

If, for fixed $\widehat{c}_L$, we solve the equation:

$$P\left(T_{n-1,\widehat{\delta}} \le 3\sqrt{n}\widehat{c}_L\right) = \gamma$$

for $\widehat{\delta}$, then we get the following expression for the observed value $\hat{b}_L$ of $\widehat{B}_L$:

$$\widehat{b}_L = k^{-1}\left(\widehat{c}_L\right) = \frac{\widehat{\delta}}{3\sqrt{n}} = \texttt{del.nct}(3 * \texttt{sqrt}(\texttt{n}) * \texttt{cL.hat}, \texttt{gam}, \texttt{n} - 1)/(3 * \texttt{sqrt}(\texttt{n}))\ ,$$

where $\texttt{gam}= \gamma$ and $\texttt{cL.hat}= \hat{c}_L$.

### 8.2 Lower Confidence Bounds for $C_U$

In a similar fashion we develop lower confidence bounds for

$$C_U = \frac{x_U - \mu}{3\sigma}\ , \qquad \text{using its natural estimate} \qquad \widehat{C}_U = \frac{x_U - \bar{X}}{3S}\ .$$

Note that in similar fashion as before we have

$$P\left(\widehat{C}_U \le k\right) = P\left(\frac{x_U - \bar{X}}{3S} \le k\right) = P\left(T_{n-1,3\sqrt{n}C_U} \le 3\sqrt{n}k\right).$$

We define $k = k(C_U)$ as that unique number which for given $C_U$ solves

$$P\left(\widehat{C}_U \le k(C_U)\right) = P\left(T_{n-1,3\sqrt{n}C_U} \le 3\sqrt{n}k(C_U)\right) = \gamma$$

As before it follows that $\widehat{B}_U = k^{-1}(\widehat{C}_U)$ serves as $100\gamma\%$ lower confidence bound for $C_U$. For an observed value $\widehat{c}_U$ of $\widehat{C}_U$ we compute the observed value $\widehat{b}_U$ of $\widehat{B}_U$ as $\widehat{\delta}/(3\sqrt{n})$, where $\widehat{\delta}$ solves

$$P\left(T_{n-1,\widehat{\delta}} \le 3\sqrt{n}\ \widehat{c}_U\right) = \gamma\ .$$

$$\text{or} \quad \widehat{b}_U = k^{-1}\left(\widehat{c}_U\right) = \frac{\widehat{\delta}}{3\sqrt{n}} = \texttt{del.nct}(3 * \texttt{sqrt}(\texttt{n}) * \texttt{cU.hat}, \texttt{gam}, \texttt{n} - 1)/(3 * \texttt{sqrt}(\texttt{n}))\ ,$$

where $\texttt{gam}= \gamma$ and $\texttt{cU.hat}= \hat{c}_U$.

### 8.3 Lower Confidence Bounds for $C_{pk}$

Putting the bounds on $C_U$ and $C_L$ together, we can obtain (slightly conservative) confidence bounds for the two-sided statistical process control parameter

$$C_{pk} = \min\left(C_L, C_U\right)$$

by simply taking

$$\widehat{B} = \min\left(\widehat{B}_L, \widehat{B}_U\right) \ .$$

If $C_L \leq C_U$, i.e., $C_{pk} = C_L$, then

$$\begin{aligned} P\left(\min\left(\widehat{B}_L, \widehat{B}_U\right) \leq \min\left(C_L, C_U\right)\right) &= P\left(\min\left(\widehat{B}_L, \widehat{B}_U\right) \leq C_L\right) \\ &\geq P\left(\widehat{B}_L \leq C_L\right) = \gamma \end{aligned}$$

and if $C_U \leq C_L$, i.e., $C_{pk} = C_U$, then

$$\begin{aligned} P\left(\min\left(\widehat{B}_L, \widehat{B}_U\right) \leq \min\left(C_L, C_U\right)\right) &= P\left(\min\left(\widehat{B}_L, \widehat{B}_U\right) \leq C_U\right) \\ &\geq P\left(\widehat{B}_U \leq C_U\right) = \gamma \ . \end{aligned}$$

Hence $\widehat{B}$ can be taken as lower bound for $C_{pk}$ with confidence level at least $\gamma$. The exact confidence level of $\widehat{B}$ is somewhat higher than $\gamma$ for $C_L = C_U$, i.e., when $\mu$ is the midpoint of the specification interval. As $\mu$ moves away from this midpoint and as $\sigma$ reduces correspondingly in order to maintain a constant $C_{pk}$ then the actual confidence level of $\widehat{B}$ gets arbitrarily close to $\gamma$ so that the confidence coefficient of $\hat{B}$ is indeed $\gamma$.

In dealing with suppliers of parts or materials one may opt for the following kind of table to communicate the issues of sampling variation and sample size. The supplier may well have some understanding about the meaning of $C_{pk}$ but it becomes somewhat hazy when $C_{pk}$ is estimated from limited data via $\hat{C}_{pk}$. Tables 1 and 2 tabulate what $\hat{C}_{pk}$ would be required in order for the $C_{pk}$ lower bound $\hat{B}$ to come out at the desired value, given in the top row of that table.

For small sample sizes $n$ the margin by which $\hat{C}_{pk}$ would have to exceed $\hat{B} = 1$ would need to be quite large. For example, when $n = 20$ we would need to have $\hat{C}_{pk} = 1.298$ in order to be 95% confident that the actual $C_{pk} \geq 1$, i.e., we would need to demonstrate that the estimated $\hat{C}_{pk}$ is .298 higher than what is desired. As $n$ gets larger, say $n = 60$ this margin can be pushed down to .150, about half of .298.

This should easily bring home the message that it pays to have a larger sample. For large $n$ the observed $\hat{C}_{pk}$ would not have to be that much larger than the desired $C_{pk}$. Of course, larger sample sizes do not guarantee better quality. If the quality is poor we are likely to see small values of $\hat{B}$ or even $\hat{C}_{pk}$, i.e., below 1. This only becomes more apparent when the sample size becomes larger.

| $n$ | desired $C_{pk}$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2 |
| 10 | 1.499 | 1.644 | 1.789 | 1.934 | 2.079 | 2.225 | 2.370 | 2.516 | 2.662 | 2.808 | 2.954 |
| 12 | 1.432 | 1.570 | 1.708 | 1.847 | 1.986 | 2.125 | 2.265 | 2.404 | 2.544 | 2.683 | 2.823 |
| 14 | 1.384 | 1.518 | 1.652 | 1.786 | 1.921 | 2.056 | 2.190 | 2.325 | 2.460 | 2.596 | 2.731 |
| 16 | 1.349 | 1.479 | 1.610 | 1.741 | 1.872 | 2.004 | 2.135 | 2.267 | 2.398 | 2.530 | 2.662 |
| 18 | 1.321 | 1.449 | 1.577 | 1.706 | 1.834 | 1.963 | 2.092 | 2.221 | 2.350 | 2.479 | 2.608 |
| 20 | 1.298 | 1.424 | 1.551 | 1.677 | 1.804 | 1.930 | 2.057 | 2.184 | 2.311 | 2.438 | 2.565 |
| 25 | 1.257 | 1.379 | 1.502 | 1.625 | 1.747 | 1.870 | 1.993 | 2.116 | 2.239 | 2.363 | 2.486 |
| 30 | 1.229 | 1.348 | 1.468 | 1.588 | 1.709 | 1.829 | 1.949 | 2.070 | 2.190 | 2.311 | 2.432 |
| 35 | 1.208 | 1.325 | 1.443 | 1.562 | 1.680 | 1.798 | 1.917 | 2.035 | 2.154 | 2.273 | 2.391 |
| 40 | 1.191 | 1.308 | 1.424 | 1.541 | 1.658 | 1.775 | 1.892 | 2.009 | 2.126 | 2.243 | 2.365 |
| 45 | 1.178 | 1.293 | 1.409 | 1.524 | 1.640 | 1.756 | 1.871 | 1.987 | 2.103 | 2.223 | 2.339 |
| 50 | 1.167 | 1.281 | 1.396 | 1.510 | 1.625 | 1.740 | 1.855 | 1.969 | 2.088 | 2.203 | 2.318 |
| 60 | 1.150 | 1.263 | 1.376 | 1.489 | 1.602 | 1.715 | 1.828 | 1.944 | 2.058 | 2.171 | 2.285 |
| 70 | 1.137 | 1.249 | 1.361 | 1.473 | 1.585 | 1.698 | 1.811 | 1.923 | 2.035 | 2.148 | 2.260 |
| 80 | 1.127 | 1.238 | 1.349 | 1.460 | 1.571 | 1.683 | 1.795 | 1.906 | 2.018 | 2.129 | 2.240 |
| 90 | 1.119 | 1.229 | 1.339 | 1.449 | 1.561 | 1.671 | 1.782 | 1.893 | 2.003 | 2.114 | 2.225 |
| 100 | 1.112 | 1.222 | 1.331 | 1.442 | 1.552 | 1.661 | 1.771 | 1.881 | 1.991 | 2.101 | 2.211 |
| 120 | 1.101 | 1.210 | 1.319 | 1.428 | 1.537 | 1.646 | 1.755 | 1.864 | 1.973 | 2.082 | 2.191 |
| 140 | 1.093 | 1.201 | 1.309 | 1.417 | 1.525 | 1.634 | 1.742 | 1.850 | 1.958 | 2.067 | 2.175 |
| 160 | 1.087 | 1.194 | 1.301 | 1.409 | 1.516 | 1.624 | 1.732 | 1.839 | 1.947 | 2.055 | 2.162 |
| 180 | 1.081 | 1.188 | 1.295 | 1.402 | 1.509 | 1.616 | 1.723 | 1.831 | 1.938 | 2.045 | 2.152 |
| 200 | 1.077 | 1.183 | 1.290 | 1.396 | 1.503 | 1.610 | 1.716 | 1.823 | 1.930 | 2.037 | 2.144 |
| 250 | 1.068 | 1.174 | 1.280 | 1.385 | 1.491 | 1.597 | 1.703 | 1.809 | 1.915 | 2.021 | 2.127 |
| 300 | 1.062 | 1.167 | 1.272 | 1.377 | 1.483 | 1.588 | 1.694 | 1.799 | 1.904 | 2.010 | 2.115 |
| 350 | 1.057 | 1.162 | 1.266 | 1.371 | 1.476 | 1.581 | 1.686 | 1.791 | 1.896 | 2.001 | 2.106 |
| 400 | 1.053 | 1.157 | 1.262 | 1.366 | 1.471 | 1.576 | 1.680 | 1.785 | 1.890 | 1.994 | 2.099 |
| 450 | 1.050 | 1.154 | 1.258 | 1.362 | 1.467 | 1.571 | 1.675 | 1.780 | 1.884 | 1.988 | 2.093 |
| 500 | 1.047 | 1.151 | 1.255 | 1.359 | 1.463 | 1.567 | 1.671 | 1.775 | 1.880 | 1.984 | 2.088 |

Table 1: Tabulated Required $\hat{C}_{pk}$ to Get as 90% Lower Bound the Desired Value of $C_{pk}$

See `http://www.boeing.com/companyoffices/doingbiz/supplier/d1-9000-1.pdf` on page 196 for a version of this table in Boeing's AQS D1-9000-1 Advanced Quality Systems Tools document for suppliers.

| | desired $C_{pk}$ | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| $n$ | 1 | 1.1 | 1.2 | 1.3 | 1.4 | 1.5 | 1.6 | 1.7 | 1.8 | 1.9 | 2 |
| 10 | 1.686 | 1.847 | 2.009 | 2.171 | 2.333 | 2.496 | 2.659 | 2.822 | 2.985 | 3.148 | 3.312 |
| 12 | 1.588 | 1.740 | 1.892 | 2.045 | 2.198 | 2.351 | 2.505 | 2.659 | 2.812 | 2.966 | 3.121 |
| 14 | 1.520 | 1.665 | 1.811 | 1.958 | 2.105 | 2.252 | 2.399 | 2.546 | 2.693 | 2.841 | 2.989 |
| 16 | 1.469 | 1.610 | 1.752 | 1.894 | 2.036 | 2.178 | 2.320 | 2.463 | 2.606 | 2.748 | 2.891 |
| 18 | 1.430 | 1.568 | 1.706 | 1.844 | 1.982 | 2.121 | 2.260 | 2.399 | 2.538 | 2.677 | 2.816 |
| 20 | 1.399 | 1.534 | 1.669 | 1.804 | 1.939 | 2.075 | 2.211 | 2.347 | 2.483 | 2.619 | 2.756 |
| 25 | 1.342 | 1.471 | 1.601 | 1.731 | 1.862 | 1.992 | 2.123 | 2.253 | 2.384 | 2.515 | 2.646 |
| 30 | 1.303 | 1.429 | 1.555 | 1.682 | 1.809 | 1.935 | 2.062 | 2.190 | 2.317 | 2.444 | 2.571 |
| 35 | 1.274 | 1.398 | 1.521 | 1.645 | 1.770 | 1.894 | 2.018 | 2.143 | 2.267 | 2.392 | 2.517 |
| 40 | 1.252 | 1.373 | 1.495 | 1.617 | 1.739 | 1.862 | 1.984 | 2.107 | 2.229 | 2.352 | 2.487 |
| 45 | 1.234 | 1.354 | 1.474 | 1.595 | 1.715 | 1.836 | 1.957 | 2.078 | 2.198 | 2.330 | 2.451 |
| 50 | 1.220 | 1.338 | 1.457 | 1.576 | 1.695 | 1.815 | 1.934 | 2.054 | 2.182 | 2.302 | 2.422 |
| 60 | 1.197 | 1.313 | 1.430 | 1.547 | 1.665 | 1.782 | 1.899 | 2.023 | 2.141 | 2.259 | 2.377 |
| 70 | 1.180 | 1.295 | 1.410 | 1.526 | 1.641 | 1.762 | 1.878 | 1.994 | 2.110 | 2.227 | 2.343 |
| 80 | 1.166 | 1.280 | 1.394 | 1.509 | 1.623 | 1.741 | 1.856 | 1.971 | 2.086 | 2.201 | 2.317 |
| 90 | 1.155 | 1.268 | 1.382 | 1.495 | 1.611 | 1.725 | 1.839 | 1.953 | 2.067 | 2.181 | 2.295 |
| 100 | 1.146 | 1.258 | 1.371 | 1.486 | 1.599 | 1.712 | 1.825 | 1.938 | 2.051 | 2.164 | 2.278 |
| 120 | 1.132 | 1.243 | 1.356 | 1.467 | 1.579 | 1.691 | 1.802 | 1.914 | 2.026 | 2.138 | 2.250 |
| 140 | 1.121 | 1.232 | 1.343 | 1.453 | 1.564 | 1.675 | 1.785 | 1.896 | 2.007 | 2.118 | 2.229 |
| 160 | 1.113 | 1.223 | 1.332 | 1.442 | 1.552 | 1.662 | 1.772 | 1.882 | 1.992 | 2.102 | 2.212 |
| 180 | 1.106 | 1.215 | 1.324 | 1.433 | 1.542 | 1.652 | 1.761 | 1.870 | 1.980 | 2.089 | 2.199 |
| 200 | 1.100 | 1.208 | 1.317 | 1.426 | 1.534 | 1.643 | 1.752 | 1.861 | 1.969 | 2.078 | 2.187 |
| 250 | 1.088 | 1.196 | 1.303 | 1.411 | 1.519 | 1.626 | 1.734 | 1.842 | 1.950 | 2.058 | 2.166 |
| 300 | 1.080 | 1.187 | 1.294 | 1.401 | 1.507 | 1.614 | 1.721 | 1.829 | 1.936 | 2.043 | 2.150 |
| 350 | 1.074 | 1.180 | 1.286 | 1.393 | 1.499 | 1.605 | 1.712 | 1.818 | 1.925 | 2.031 | 2.138 |
| 400 | 1.069 | 1.174 | 1.280 | 1.386 | 1.492 | 1.598 | 1.704 | 1.810 | 1.916 | 2.022 | 2.128 |
| 450 | 1.064 | 1.170 | 1.275 | 1.381 | 1.486 | 1.592 | 1.698 | 1.803 | 1.909 | 2.015 | 2.120 |
| 500 | 1.061 | 1.166 | 1.271 | 1.376 | 1.482 | 1.587 | 1.692 | 1.798 | 1.903 | 2.008 | 2.114 |

Table 2: Tabulated Required $\hat{C}_{pk}$ to Get as 95% Lower Bound the Desired Value of $C_{pk}$

## 9 Coefficient of Variation Confidence Bounds

The coefficient of variation is traditionally defined as the ratio of standard deviation to mean, i.e., as $\nu = \sigma/\mu$. It expresses the amount of measurement variability relative to what is being measured. We will instead give confidence bounds for its reciprocal $\rho = 1/\nu = \mu/\sigma$. The reason for this is that $\bar{X}$, in the natural estimate $S/\bar{X}$ for $\nu$, could be near zero, causing certain problems. If the coefficient of variation is sufficiently small, usually the desired situation, then the distinction between it and its reciprocal is somewhat immaterial since typical bounds for $\nu$ can be inverted to bounds for $\rho$ and vice versa. This situation is easily recognized by the sign of the upper or lower bound, respectively. If $\hat{\rho}_L$ as lower bound for $\rho$ is positive, then $\hat{\nu}_U = 1/\hat{\rho}_L$ is an upper bound for a positive value of $\nu$. If $\hat{\rho}_U$ as upper bound for $\rho$ is negative, then $\hat{\nu}_L = 1/\hat{\rho}_U$ is a lower bound for a negative value of $\nu$. In either case $\rho$ is bounded away from zero which implies that the reciprocal $\nu = 1/\rho$ is bounded. On the other hand, if $\hat{\rho}_L$ as lower bound for $\rho$ is negative, then $\rho$ is not bounded away from zero and the reciprocal values could be arbitrarily large. Hence in that case $\hat{\nu}_U = 1/\hat{\rho}_L$ is useless as an upper bound for $\nu$ since no finite upper bound on the values of $\nu$ can be derived from $\hat{\rho}_L$.
To construct a lower confidence bound for $\rho = \mu/\sigma$ consider

$$\sqrt{n}\frac{\bar{X}}{S} = \frac{\sqrt{n}(\bar{X} - \mu)/\sigma + \sqrt{n}\mu/\sigma}{S/\sigma} = T_{n-1,\,\delta}$$

with $\delta = \sqrt{n}\mu/\sigma$. Again the random variable

$$U = G_{n-1,\,\delta}(\sqrt{n}\ \bar{X}/S) = G_{n-1,\,\delta}(T_{n-1,\,\delta})$$

is distributed uniformly over $(0, 1)$. Hence $P(U \le \gamma) = \gamma$ so that

$$G_{n-1,\,\delta}(\sqrt{n}\ \bar{X}/S) \le \gamma \quad \text{if and only if} \quad \hat{\delta}_L \le \delta\ ,$$

where $\hat{\delta}_L$ is the solution of

$$G_{n-1,\,\hat{\delta}_L}(\sqrt{n}\ \bar{X}/S) = \gamma \tag{10}$$

and $\hat{\rho}_L \stackrel{\text{def}}{=} \hat{\delta}_L/\sqrt{n} = \texttt{del.nct}(\texttt{sqrt(n)} * \texttt{Xbar/S}, \texttt{gam}, \texttt{n} - \texttt{1})/\texttt{sqrt(n)}$ is thus a $100\gamma\%$ lower confidence bound for $\rho = \delta/\sqrt{n} = \mu/\sigma$. Here $\texttt{Xbar} = \bar{X}$ and $\texttt{gam} = \gamma$.
To obtain an upper bound for $\rho$ with confidence level $\gamma$ one finds $\hat{\delta}_U$ as solution of

$$G_{n-1,\,\hat{\delta}_U}(\sqrt{n}\ \bar{X}/S) = 1 - \gamma \tag{11}$$

and uses $\hat{\rho}_U \stackrel{\text{def}}{=} \hat{\delta}_U/\sqrt{n} = \texttt{del.nct}(\texttt{sqrt(n)} * \texttt{Xbar/S}, \texttt{1} - \texttt{gam}, \texttt{n} - \texttt{1})/\texttt{sqrt(n)}$ as $100\gamma\%$ upper bound for $\rho = \delta/\sqrt{n} = \mu/\sigma$. Here $\texttt{Xbar} = \bar{X}$ and $\texttt{gam} = \gamma$.

## 10 Batch Effects

Eventually these methods for tolerance bounds or confidence bounds for $C_L, C_u$, or $C_{pk}$ were also used in the context of composite materials where batch effects can be quite significant. Batch effects result from changing chemical compositions for each batch of material and a good portion of the strength variation of tested specimens from that material is due to the variation from batch to batch. Early in the production only few batches are typically available. Batches are expensive. Often many specimens from each of the few batches are tested or measured in the hope of making up for the deficit of having only few batches available.

There are two extreme situations:

1. The variation from batch to batch is insignificant and one can treat all of the specimen strengths as one big sample of size $N = n_1 + \ldots + n_k$, where $k$ is the number of batches involved and $n_i$ is the number of strength measurements from the $i^{\text{th}}$ batch.

2. The variation from batch to batch is so strong compared to the variation within batches, that it is a wasted effort to have more than one observation per batch. Having $n_i > 1$ only serves the purpose of ascertaining that variability mismatch. Taking $n_i$ observations from the $i^{\text{th}}$ batch in that case is like writing down the "same" test result $n_i$ times. Treating all $N = n_1 + \ldots + n_k$ as one large random sample greatly inflates the "effective" sample size. To be more realistic, we should just work with one observation per batch and let the batch to batch variation speak for itself. In that case the real "effective sample size" should be $k$.

This problem was addressed by Scholz and Vangel (1998) by interpolating between these two extreme situations and reducing the problem to that of a simple random sample of some "effective" sample size $N^\star$ somewhere between $k$ and $N$, that reflects the ratio of within to between batch variability. This reduced a rather messy situation in a simple and intuitive fashion to the previous process for a pure random sample.

The same process gave solutions for tolerance bounds and confidence bounds for the capability indices $C_L, C_u$, or $C_{pk}$. We will here address only tolerance bounds and refer to the above reference for the other situation. Presumably the process would also carry over to confidence bounds for tail probabilities due to the duality with tolerance bounds but we have not checked the details.

We used the following measurement variation model $X_{ij} = \mu + b_i + e_{ij}$, $j = 1, \ldots, n_i$ and $i = 1, \ldots, k$, where $b_i$ (between batch variation effect) is normal with mean zero and variance $\sigma_b^2$ and $e_{ij}$ (within batch variation effects) is normal with mean zero and variance $\sigma_e^2$. The effects $b_i$ and $\{e_{ij}\}$ are assumed to be mutually independent. Hence $X_{ij} \sim \mathcal{N}(\mu, \sigma_b^2 + \sigma_e^2)$ . The correlation of two different observations within the same batch is $\rho = \sigma_b^2/(\sigma_b^2 + \sigma_e^2)$ which can range anywhere within $[0, 1]$.

The individual sample sizes $n_i$ from each batch can vary. However, in developing the ultimate solution we were guided strongly by the special case $n_1 = \ldots = n_k$. Even in that case we invoked

an interpolation approximation. This was augmented with a further approximation (Satterthwaite) when allowing the $n_i$ to be different. Simulations were then used to check whether these approximations gave reasonable results.

This solution is an typical example of industrial statistics, where a quick fix to a messy problem was required. It arose when a supplier was trying to build his case based on one large sample $N = n_1 + \ldots + n_k$ without accounting for the possible batch effects. After confirming the significance of that effect it was essential to find a middle ground, which was easily captured by the "effective sample size" concept, since it reduced the calculations in a simple manner to a previously accepted method.

The main idea is to conceptualize a pure random sample $X_1^\star, \ldots, X_{N^\star}^\star$ from $\mathcal{N}(\mu, \sigma_b^2 + \sigma_e^2)$ that carries the "same kind of information" as the original data. $N^\star$ then represents the "equivalent sample size."

Since $\bar{X} = \sum_{i=1}^{B} \sum_{j=1}^{n_i} X_{ij}/N$ and $\bar{X}^\star = \sum_{i=1}^{N^\star} X_i^\star/N^\star$ both are normally distributed with same mean $\mu$ we implement the notion of "same kind of information" by choosing $N^\star$ to match the variances of $\bar{X}$ and $\bar{X}^\star$, i.e., find $N^\star$ such that

$$\operatorname{var}\left(\bar{X}\right) = \operatorname{var}\left(\mu + \frac{\sum_{i=1}^{k} n_i b_i + \sum_{i=1}^{k} \sum_{j=1}^{n_i} e_{ij}}{N}\right) = \sigma_b^2 \sum_{i=1}^{k} \left(\frac{n_i}{N}\right)^2 + \sigma_e^2 \frac{1}{N} = \operatorname{var}\left(\bar{X}^\star\right) = \frac{\sigma_b^2 + \sigma_e^2}{N^\star} .$$

This leads to the following formula for $N^\star = N^\star(\rho)$

$$N^\star = \left[\frac{\sigma_b^2}{\sigma_b^2 + \sigma_e^2} \sum_{i=1}^{k} \left(\frac{n_i}{N}\right)^2 + \frac{1}{N} \frac{\sigma_e^2}{\sigma_b^2 + \sigma_e^2}\right]^{-1} = \left[\rho \frac{1}{f+1} + (1-\rho) \frac{1}{N}\right]^{-1} ,$$

where we write $1/(f+1) = \sum_{i=1}^{k} (n_i/N)^2$ for reasons to become clear later. Note that $N^\star$ is the weighted harmonic mean of $f+1$ and $N$.

For $\rho = 0$ this becomes $N^\star = N$ and for $\rho = 1$ we get $N^\star = f + 1$ which matches $k$ when $n_1 = \ldots = n_k$. Thus in the latter case of equal batch sizes this effective sample size formula agrees with our previous notion of what the effective sample size should be in these two extreme situations.

We will not bother with the fact that $N^\star$ may not be an integer. An actual conceptual sample $X_1^\star, \ldots, X_{N^\star}^\star$ is never used in our procedure and all calculations are based on the original batch data $\{X_{ij}\}$.

In practice the within batch correlation $\rho$ is unknown but one may find reasonable estimates from the data as follows. Compute the between batch and error sums of squares

$$SS_b = \sum_{i=1}^{k} n_i (\bar{X}_{i\cdot} - \bar{X})^2 \qquad \text{and} \qquad SS_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (X_{ij} - \bar{X}_{i\cdot})^2 .$$

Take $\hat{\sigma}_e^2 = SS_e/(N-k)$ as unbiased estimate of $\sigma_e^2$ and $\hat{\tau}^2 = SS_b/(k-1)$ as unbiased estimate of

$$\tau^2 = \sigma_e^2 + \sigma_b^2 \, \frac{N}{k-1} \left(1 - \sum_{i=1}^{k} \left(\frac{n_i}{N}\right)^2\right) = \sigma_e^2 + \sigma_b^2 \, \frac{N}{k-1} \, \frac{f}{f+1} \, .$$

Combining these two estimates we get $\hat{\sigma}_b^2 = (\hat{\tau}^2 - \hat{\sigma}_e^2)\,(k-1)(f+1)/(N\,f)$ as unbiased estimate for $\sigma_b^2$. Unfortunately, this latter estimate may be negative. If that happens it is suggested to set the estimate to zero. We denote this modification again by $\hat{\sigma}_b^2$ but it will no longer be unbiased. The estimate of $\rho$ is then computed as $\hat{\rho} = \hat{\sigma}_b^2/(\hat{\sigma}_b^2 + \hat{\sigma}_e^2)$. It is this estimate that is used in place of $\rho$ in estimating $N^\star$ by $\hat{N}^\star = N^\star(\hat{\rho})$.
We will now focus on tolerance bounds under the two previously discussed extreme scenarios: no batch to batch variation and no within batch variation.

### 10.1 No Between Batch Variation

Here we assume $\sigma_b = 0$ and $\sigma_e > 0$, i.e, $\rho = 0$, and thus all observations $X_{ij}$ are mutually independent. $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/N)$ and $SS_T = SS_b + SS_e = (N-1)S^2 \sim \sigma^2 \cdot \chi^2_{N-1}$ and both are independent of each other.
In Section 6 it was shown that $100\gamma\%$ lower tolerance bounds are of the form $\bar{X} - k\,S$, where $k$

$$k = k_0(N) = \frac{1}{\sqrt{N}} \, t_{N-1,-z_p\sqrt{N},\gamma} = \sqrt{\frac{N-1}{N}} \, \frac{1}{\sqrt{N-1}} \, t_{N-1,-z_p\sqrt{N},\gamma} \, , \tag{12}$$

where $t_{N-1,-z_p\sqrt{N},\gamma}$ is the $\gamma$-quantile of $T_{N-1,-z_p\sqrt{N}}$.

### 10.2 No Within Batch Variation

Here we assume $\sigma_b > 0$ and $\sigma_e = 0$, i.e, $\rho = 1$, and thus $\sigma^2 = \sigma_b^2$ and all observations within each batch are identical. Hence $SS_e = 0$, and thus $S^2 = SS_b/(N-1)$. Using Satterthwaite's method we will approximate the distribution of $SS_T = SS_b$ by a chi-square multiple with $g$ degrees of freedom, i.e., $SS_T = SS_b \approx a \cdot \chi^2_g$, where $a$ and $g$ are determined to match the first two moments or mean and variance on either side. As shown in Appendix B this leads to

$$g = \frac{(1 - \sum w_i^2)^2}{\sum w_i^2 - 2\sum w_i^3 + (\sum w_i^2)^2} \quad \text{and} \quad a = \frac{N}{g} \, \sigma_b^2 \, \left(1 - \sum_{i=1}^{k} w_i^2\right) = \frac{N}{g} \, \sigma_b^2 \, \frac{f}{f+1} \, ,$$

where $w_i = n_i/N$ and summations are over $i = 1, \ldots, k$. In Appendix C it is shown that this complicated expression for $g$ can be approximated very well by a much simpler expression, namely by $f = (\sum w_i^2)^{-1} - 1$, and the Satterthwaite approximation is exact when the $n_i$ are all the same. We

will use this simplification ($f$ replacing $g$) from now on since it leads to a convenient similarity of the formulas for the factor $k$ in the two cases studied. With this simplification we have $a \approx N\,\sigma_b^2/(f+1)$ and we can treat

$$V^2 = \frac{SS_T}{a\ f} = \frac{SS_b}{a\ f} = S^2\ \frac{(N-1)(f+1)}{f\ N\ \sigma_b^2}$$

as an approximate $\chi_f^2/f$ random variable. Further, $\bar{X} \sim \mathcal{N}(\mu, \tau^2)$ with $\tau^2 = \sigma_b^2 \cdot \sum_{i=1}^k w_i^2 = \sigma_b^2/(f+1)$, i.e., $Z = \sqrt{f+1}\,(\bar{X}-\mu)/\sigma_b$ has a standard normal distribution.

Note that when all samples sizes $n_i$ are the same ($= n$), then the above complicated expressions for $f$ and $a$ (and their approximations) reduce to $f = k-1$ and $a = n\sigma_b^2$. In that case $SS_b$ actually is exactly distributed like $n\sigma_b^2 \cdot \chi_{k-1}^2$ and then $SS_T = SS_b$ is independent of $\bar{X}$. When the samples sizes are not the same, then $SS_T$ is approximately distributed like the above chi-square multiple and the strict independence property no longer holds. We will ignore this latter flaw in our derivation below. The simulations show that this is of no serious consequence.

Again we have

$$\begin{aligned}
\gamma = P\left(\bar{X} - k\ S \le x_p\right) &= P\left(\frac{\sqrt{f+1}\,(\bar{X}-\mu)}{\sigma_b} - \frac{\sqrt{f+1}\,(x_p-\mu)}{\sigma_b} \le \frac{k\sqrt{f+1}\,S}{\sigma_b}\right) \\
&= P\left(\frac{Z - z_p\sqrt{f+1}}{V} \le k\sqrt{\frac{f\ N}{N-1}}\ \right) \\
&= P\left(T_{f,-z_p\sqrt{f+1}} \le k\sqrt{\frac{f\ N}{N-1}}\ \right)
\end{aligned}$$

leading to

$$k = k_1(N) = \sqrt{\frac{N-1}{N}}\ \frac{1}{\sqrt{f}}\ t_{f,-z_p\sqrt{f+1},\gamma}\ . \tag{13}$$

We note the strong parallelism between equations (12) and (13) for the $k$-factor. Aside from the common factor $\sqrt{(N-1)/N}$ in both expressions, the expressions match in the sense of using the respective effective sample size $N$ and $f+1$. Note that the actual tolerance bound is of the form $\bar{X} - kS$ in both these extreme cases.

### 10.3 The Interpolation Step

We note that the two expressions for $k_0(N)$ and $k_1(N)$ in equations (12) and (13) share the common factor $\sqrt{(N-1)/N}$ and the remainder can be matched if we interchange $f+1$ and $N$. The actual tolerance bound is of the form $\bar{X} - kS$ in both these extreme cases.

For batch effect situations that are positioned between these two extreme cases we propose to use the previously developed estimated effective sample size $\hat{N}^\star$ as a simple interpolation between $f+1$ and $N$ and use as $k$-factor in the general case

$$k^\star(N) = \sqrt{\frac{N-1}{N}}\ \frac{1}{\sqrt{\hat{N}^\star - 1}}\ t_{\hat{N}^\star - 1, -z_p\sqrt{\hat{N}^\star}, \gamma}\ .$$

### 10.4 An Example Calculation

The data in Table 3 represent data on 21 batches of some composite material property data. From the data in this table we obtain: $\bar{X} = 49.638$ and $S = 1.320$. Ignoring the batch effects and assuming that we deal with $N = 63$ independent observations we obtain as $k$-factor for the $A$-allowable

$$k_A = \texttt{qnct}(.95, 63-1, -\texttt{qnorm}(.01) * \texttt{sqrt}(63))/\texttt{sqrt}(63) = 2.793392$$

and thus $A = \bar{X} - k_A S = 49.638 - 2.793392 * 1.320 = 45.95072$ as $A$-allowable.
However, the given data show strong batch effects, see Figure 19, and the above allowable may not be appropriate. When adjusting by the "effective" sample size we obtain

$$SS_b = 78.921\ , \quad SS_e = 29.148\ , \quad f = 17.123\ , \quad \hat{\sigma}_e^2 = .6939\ , \quad \hat{\sigma}_b^2 = 1.093$$

and thus $\hat{\rho} = .6116$ and $N^\star = 25.056$. As $k$-factor for the $A$ allowable we now get

$$\begin{aligned} k_A &= \texttt{sqrt}((63-1)/63) * \texttt{qnct}(.95, 25.056-1, -\texttt{qnorm}(.01) * \texttt{sqrt}(25.056))/\texttt{sqrt}(25.056-1) \\ &= 3.195986 \end{aligned}$$

and thus $A = \bar{X} - k_A S = 49.638 - 3.195986 * 1.320 = 45.4193$ as $A$-allowable.
If the threshold, against which these allowables are compared, had been 45 then the allowables by either analysis fall on the same side of 45, namely above. However, if the threshold had been 45.5 then the allowables fall on opposite sides of 45.5, the one accounting for the batch effect falling a little bit short. This may be mainly because of the "effective" sample size being too small.
A closer examination of Figure 19 suggests that the measured values stabilize from batch 14 onward. Prior to that point the batch to batch variation seems quite strong. Also, there may have been selective decisions on how many data points to gather, depending on what was seen on the first and/or second measurement in each batch. Such a selection bias would put in doubt any of the calculations made so far. There is no realistic way to account for such bias.

Table 3: Example Batch Data

| batch | $n_i$ | sample data | $\bar{X}_{i\cdot}$ |
|---|---|---|---|
| 1 | 1 | 50.5 | 50.5 |
| 2 | 1 | 50.2 | 50.2 |
| 3 | 4 | 50.7, 50.8, 51.4, 51.3 | 51.05 |
| 4 | 1 | 49.3 | 49.3 |
| 5 | 3 | 51.0, 51.2, 53.4 | 51.867 |
| 6 | 3 | 50.9, 51.6, 51.8 | 51.433 |
| 7 | 1 | 49.3 | 49.3 |
| 8 | 3 | 48.6, 48.2, 46.6 | 47.8 |
| 9 | 2 | 50.4, 49.9 | 50.15 |
| 10 | 2 | 48.2, 47.5 | 47.85 |
| 11 | 3 | 50.5, 48.2, 49.5 | 49.4 |
| 12 | 3 | 49.7, 51.4, 50.6 | 50.567 |
| 13 | 4 | 49.6, 51.1, 51.1, 52.5 | 51.075 |
| 14 | 4 | 48.4, 50.2, 48.8, 49.1 | 49.125 |
| 15 | 4 | 48.8, 49.8, 50.0, 50.5 | 49.775 |
| 16 | 5 | 49.3, 50.2, 49.8, 48.9, 48.7 | 49.38 |
| 17 | 4 | 49.3, 47.5, 49.4, 48.4 | 48.65 |
| 18 | 4 | 47.8, 47.7, 48.8, 49.9 | 48.55 |
| 19 | 3 | 50.0, 49.5, 49.3 | 49.6 |
| 20 | 4 | 48.5, 49.2, 48.3, 47.8 | 48.45 |
| 21 | 4 | 47.9, 49.6, 49.8, 49.0 | 49.075 |

If we disregard these first 13 batches and obtain an $A$-allowable from the remaining 8 batches with a total of 32 observations we find $\bar{X} = 49.06875$ (not much changed) and $S = 0.8133711$ (quite a bit smaller) and the $k$-factor becomes

$$\texttt{qnct}(.95, 32-1, -\texttt{qnorm}(.01)*\texttt{sqrt}(32))/\texttt{sqrt}(32) = 3.033847$$

with resulting $A$-allowable $A =$`49.06875-3.033847*0.8133711=46.60111`.
Using the above interpolation method we find $N^\star = 22.44343$, $k_A = 3.243241$ and $A = 46.43079$, which is not that much different from 46.60111 and both values are significantly higher than the previous ones based on the full data set.

### 10.5 Validation Simulations

Simulations of the above process were run for the corresponding bounds on $C_{pk}$. Since these were not discussed here we will not reproduce the results. For various magnitudes of batch effects we observed the coverage rate of these bounds and found that the actual coverage came close to the nominal one, if not a bit higher. On the other hand, the coverage probability of the method that ignored the batch effect fell off strongly as the batch variation became more and more dominant. For details see the reference or the posted preprint.

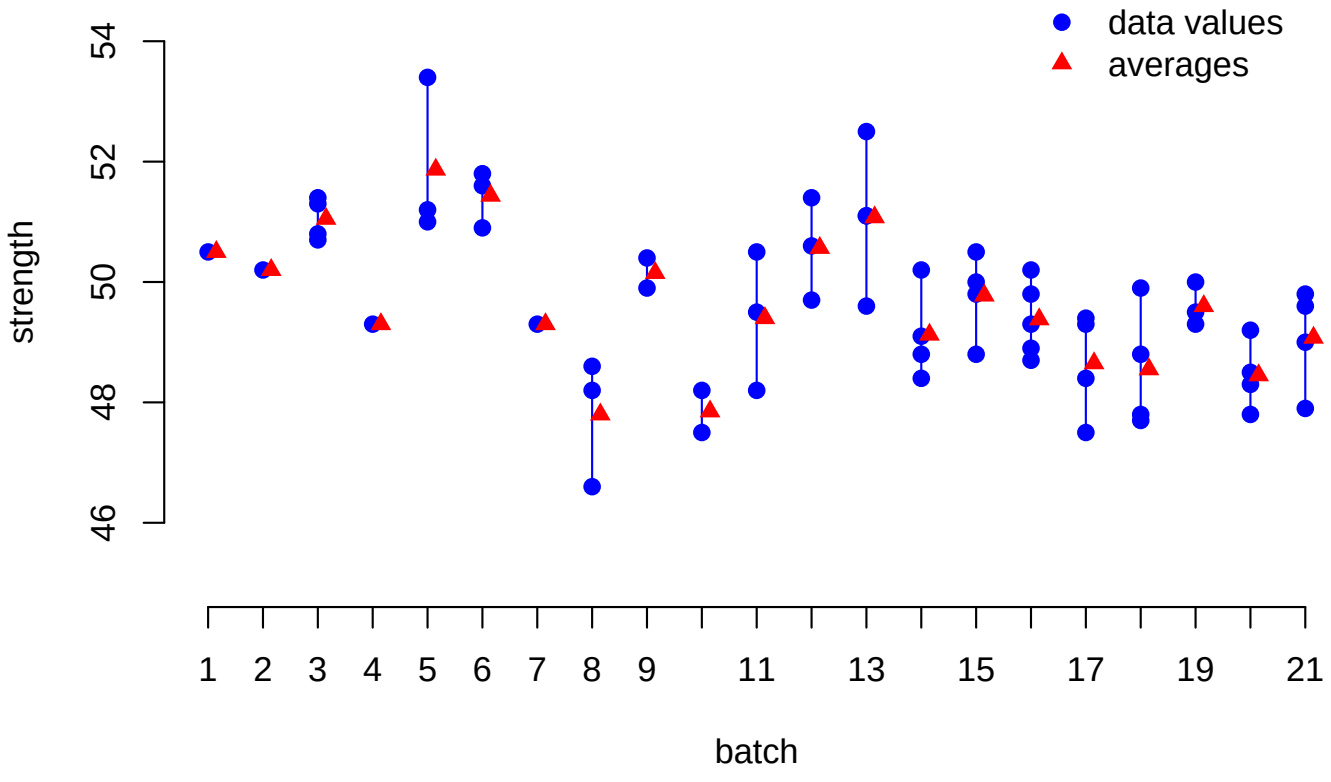



Figure 19: Batch Data

## 11 Tolerance Bounds in Regression

As indicated in the Introduction, the methodology involving applications of the noncentral $t$-distribution to single normal random samples can easily be extended to more complex data situations. We will show here how this is done for tolerance bounds in the context of regression.

The standard linear regression model assumes the following data structure for $n$ observations or responses $Y_1, \ldots, Y_n$, observed under respectively varying but known conditions $\boldsymbol{x}_1' = (x_{11}, \ldots, x_{1p})$, $\ldots$, $\boldsymbol{x}_n' = (x_{n1}, \ldots, x_{np})$

$$Y_i = x_{i1}\beta_1 + \ldots + x_{ip}\beta_p + e_i = \boldsymbol{x}_i'\boldsymbol{\beta} + e_i \ , \qquad i = 1, \ldots, n \ .$$

Here $\beta_1, \ldots, \beta_p$ are unknown parameters that are to be estimated from the data $(Y_1, \boldsymbol{x}_1'), \ldots, (Y_n, \boldsymbol{x}_n')$. The terms $e_1, \ldots, e_n$ are the error terms that capture to what extent the observed values $Y_i$ differ from the model values $\boldsymbol{x}_i'\boldsymbol{\beta}$. It is typically assumed that these error terms are statistically independent with common $\mathcal{N}(0, \sigma^2)$ distribution, where the variance $\sigma^2$ is also unknown, to be estimated from the data as well.

As a concrete example we consider the tensile strength of coupons of composite materials. These consist of laminates, i.e., are built up from layers of lamina, typically using lamina with varying fiber ply orientations, such as 90°, 45° and 0°. Such laminates are usually characterized by the percent of lamina in each orientation. Since these percentages have to add up to 100% it is only necessary to specify $k - 1 = 2$ percentages when $k = 3$ orientations are involved. Here the response $Y$ is the tensile strength of the coupon (the force at which it breaks under tension) and $\boldsymbol{x} = (x_1, x_2)$ gives the two percentages for lamina at 45° and 0° orientation. In addition to the simple linear model in the covariates $(x_1, x_2)$ one may also want to explore any quadratic effects, i.e., $x_3 = x_1^2, x_4 = x_2^2, x_5 = x_1x_2$.

Testing such coupons is costly. Since there are many possible lay-up orientation combinations, it becomes prohibitive to test all these combinations extensively. Thus it makes sense to test coupons in moderate numbers for several such combinations, carefully chosen to cover the space of lay-up percentages reasonably well. Upfront it is not known which lay-up combination will give the best strength results and it is entirely possible that coupons at such an optimal combination have not been tested at all for the initial experiment. However, such test runs can be added later in confirmatory testing or in order to tighten up the tolerance bounds.

The full data set would then consist of $(Y_1, x_{11}, x_{21}), \ldots, (Y_n, x_{1n}, x_{2n})$. If the quadratic model is entertained this expands to $(Y_1, x_{11}, \ldots, x_{51}), \ldots, (Y_n, x_{1n}, \ldots, x_{5n})$. Much of the variation in the strength measurement $Y$ comes from testing itself. Both the orientation at which the stress is applied and the orientation of the coupon as it is cut from the manufactured laminate can vary and thus could have significant strength impact. Of course there are other factors that can cause response variation, for example chemical batch effects as mentioned previously in Section 10. Here

we will confine ourselves to the pure regression model. However, it should be possible to blend the methods of Section 10 with the solution for this pure regression model.

The above equations for the pure regression model can be written more concisely in terms of matrix notation

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11}\beta_1 + \ldots + x_{1p}\beta_p \\ x_{21}\beta_1 + \ldots + x_{2p}\beta_p \\ \vdots \\ x_{n1}\beta_1 + \ldots + x_{np}\beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix}$$

$$= \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1p} \\ x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}\ .$$

It is usually assumed that $n > p$ and that the matrix $\boldsymbol{X}$ is of full rank $p$, i.e., its $p$ columns $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_p \in R^n$ are linearly independent. This means that the equation $a_1\boldsymbol{x}_1 + \ldots + a_p\boldsymbol{x}_p = \boldsymbol{0}$ only admits the solution $\boldsymbol{a}' = (a_1, \ldots, a_p) = (0, \ldots, 0)$. This entails that the $p \times p$ matrix $\boldsymbol{X}'\boldsymbol{X}$ has full rank $p$ as well[2] and thus the equation $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{a} = \boldsymbol{b}$ has a unique solution $\boldsymbol{a} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{b}$ for each $\boldsymbol{b}$. Here $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ is the inverse matrix to $\boldsymbol{X}'\boldsymbol{X}$. A $p \times p$ matrix $\boldsymbol{A}$ is the inverse to a $p \times p$ matrix $\boldsymbol{B}$ if $AB = I = I_p$, where $I_p$ is a $p \times p$ matrix with 1's on the diagonal and 0's off the diagonal.

Multiplying the above matrix version of the data model by $\boldsymbol{X}'$ and then by $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ we get

$$\boldsymbol{X}'\boldsymbol{Y} = \boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{X}'\boldsymbol{e} \implies (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} = \boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}$$

$$\hat{\boldsymbol{\beta}} = \beta + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e} \quad \text{where} \quad \hat{\boldsymbol{\beta}} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$$

is the least squares estimate of $\beta$. The name "least squares estimate" derives from the fact that this vector $\hat{\boldsymbol{\beta}}$ is the vector $\boldsymbol{\beta}$ that minimizes the following sum of squares

$$\sum_{i=1}^{n}(Y_i - \boldsymbol{x}_i'\boldsymbol{\beta})^2 = \sum_{i=1}^{n}(Y_i - x_{i1}\beta_1 - \ldots - x_{ip}\beta_p)^2$$

$$= (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) = (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}} + \boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})$$

$$= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})$$

$$+ (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})$$

$$= (\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) + (\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{X}\hat{\boldsymbol{\beta}} - \boldsymbol{X}\boldsymbol{\beta}) \tag{14}$$

[2] $\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u} = \boldsymbol{0} \implies \boldsymbol{u}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{u} = 0 \implies \boldsymbol{X}\boldsymbol{u} = \boldsymbol{0} \implies \boldsymbol{u} = \boldsymbol{0}$.

where the simplification to the two terms in the last equation derives from the fact that the two middle terms in the previous equation vanish, as is shown here only for the first of these terms since the other is just its transpose:

$$\begin{aligned}(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})'(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta}) &= (\boldsymbol{Y}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y})'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &= \boldsymbol{Y}'(\boldsymbol{X}-\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}) \\ &= \boldsymbol{Y}'(\boldsymbol{X}-\boldsymbol{X})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\boldsymbol{0}\ .\end{aligned}$$

The second term in (14) is minimized by taking $\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}$ as is seen from

$$(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{X}\hat{\boldsymbol{\beta}}-\boldsymbol{X}\boldsymbol{\beta})=(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'\boldsymbol{X}'\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\geq 0\ ,$$

with equality if and only if $\boldsymbol{X}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})=\boldsymbol{0}$, i.e., if $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=\boldsymbol{0}$. This proves the least squares property of $\hat{\boldsymbol{\beta}}$ since the other term in (14) does not depend on $\boldsymbol{\beta}$.

Suppose we want to understand the response $Y(\boldsymbol{x}_0)$ under the experimental conditions $\boldsymbol{x}_0'=(x_{01}\ldots,x_{0p})$, then $Y(\boldsymbol{x}_0)\sim\mathcal{N}(\boldsymbol{x}_0'\boldsymbol{\beta},\sigma^2)$. The natural estimate of $\boldsymbol{x}_0'\boldsymbol{\beta}$ is

$$\hat{Y}(\boldsymbol{x}_0)=\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}=\boldsymbol{x}_0'\beta+\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}\sim\mathcal{N}(\boldsymbol{x}_0'\beta,\sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0)=\mathcal{N}(\mu(\boldsymbol{x}_0),\tau^2(\boldsymbol{x}_0))\ ,$$

where the mean $\mu(\boldsymbol{x}_0)=\boldsymbol{x}_0'\beta$ derives from the fact that $E(e_i)=0$ for $i=1,\ldots,n$ and the variance expression $\tau^2(\boldsymbol{x}_0)=\sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0$ comes from

$$\begin{aligned}\mathrm{var}(\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{e}) &= \mathrm{var}(\boldsymbol{u}'\boldsymbol{e})=\sigma^2\sum_{i=1}^{n}u_i^2=\sigma^2\boldsymbol{u}'\boldsymbol{u} \\ &= \sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0=\sigma^2\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0\ .\end{aligned}$$

The unknown parameter $\sigma^2$ can be estimated by the unbiased estimator

$$S^2=\frac{1}{n-p}\sum_{i=1}^{n}(Y_i-\boldsymbol{x}_i'\hat{\boldsymbol{\beta}})^2=\frac{1}{n-p}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2=\frac{1}{n-p}\sum_{i=1}^{n}\hat{e}_i^2\ ,$$

where the $\hat{Y}_i=\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}$ and $\hat{e}_i=Y_i-\hat{Y}_i$, $i=1,\ldots,n$, are also known as fitted values and residuals, respectively. It is known that $(n-p)S^2/\sigma^2$ has a $\chi^2_{n-p}$ distribution and is independent of $\hat{\boldsymbol{\beta}}$ and thus also independent of $\hat{Y}(\boldsymbol{x}_0)=\boldsymbol{x}_0'\hat{\boldsymbol{\beta}}$.

The $p$-quantile of the response $Y(\boldsymbol{x}_0)$ is $y_p(\boldsymbol{x}_0)=\mu(\boldsymbol{x}_0)+\sigma z_p$ and its natural estimate is $\hat{Y}(\boldsymbol{x}_0)+Sz_p$. As in the case of a single random sample, where we considered tolerance bounds of the form $\bar{X}-kS$, we will now consider similarly constructed tolerance bounds in the regression situation, namely

$\hat{Y}(\boldsymbol{x}_0) - k(\boldsymbol{x}_0)S$. Note that the $k$-factor here depends on $\boldsymbol{x}_0$, the reason for which becomes clear in the derivation. From the above we have

$$Z = \frac{\hat{Y}(\boldsymbol{x}_0) - \mu(\boldsymbol{x}_0)}{\tau(\boldsymbol{x}_0)} = \frac{\hat{Y}(\boldsymbol{x}_0) - \mu(\boldsymbol{x}_0)}{\sigma\sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}} \sim \mathcal{N}(0,1) \qquad \text{and} \qquad V = \frac{S^2(n-p)}{\sigma^2} \sim \chi^2_{n-p}$$

are independent. Abbreviating $\kappa(\boldsymbol{x}_0) = \sqrt{\boldsymbol{x}_0'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_0}$ we continue with

$$\begin{aligned}
\gamma = P(\hat{Y}(\boldsymbol{x}_0) - kS \le y_p(\boldsymbol{x}_0)) &= P(\hat{Y}(\boldsymbol{x}_0) - kS \le \mu(\boldsymbol{x}_0) + \sigma z_p) \\
&= P\left(\frac{\hat{Y}(\boldsymbol{x}_0) - \mu(\boldsymbol{x}_0) - \sigma z_p}{\sigma\kappa(\boldsymbol{x}_0)} \le \frac{kS}{\sigma\kappa(\boldsymbol{x}_0)}\right) \\
&= P\left(\frac{Z - z_p/\kappa(\boldsymbol{x}_0)}{\sqrt{V/(n-p)}} \le \frac{k}{\kappa(\boldsymbol{x}_0)}\right) \\
&= P\left(T_{n-p,\delta(\boldsymbol{x}_0)} \le k/\kappa(\boldsymbol{x}_0)\right) = G_{n-p,\delta(\boldsymbol{x}_0)}(k/\kappa(\boldsymbol{x}_0)) \ ,
\end{aligned}$$

where $\delta(\boldsymbol{x}_0) = -z_p/\kappa(\boldsymbol{x}_0)$. Thus $k = \kappa(\boldsymbol{x}_0)G^{-1}_{n-p,\delta(\boldsymbol{x}_0)}(\gamma)$= `kappa` $*$ `qnct(gam, n - p, delta)`, where `delta` $=\delta(\boldsymbol{x}_0)$, `kappa` $=\kappa(\boldsymbol{x}_0)$ and `gam` $=\gamma$. The two-fold dependence of $k$ on $\boldsymbol{x}_0$ should now be quite obvious.

The R workspace contains a function `reg.tolbd` that calculates such $100\gamma\%$ lower confidence bounds for $y_p(\boldsymbol{x}_0)$ for any specified $(\gamma, p, \boldsymbol{x}_0)$. Note that the intercept covariate is assumed and is not input into this function, it is created internally. Since a $100(1-\gamma)\%$ lower bound is a $100\gamma\%$ upper bound, the same procedure can be used for getting upper bounds. The documentation to `reg.tolbd` is given in the function body.

In addition we provided a corresponding function, called `poly.tolbd`, that is tailored to polynomial fits with respect to a univariate explanatory variable. These polynomial models are special cases of the general linear regression model, as is seen from the following response model:

$$Y_i = x_{i0}\beta_0 + x_{i1}\beta_1 + \ldots + x_{ik}\beta_k + e_i \ ,$$

where the explanatory variables $x_{ij}$ have the form $x_{ij} = x_i^j,\ j = 0, 1, \ldots, k$ and the $x_1, \ldots, x_n$ are $n$ observations on a univariate covariate observed in conjunction with the responses $Y_1, \ldots, Y_n$. The special treatment of creating a separate function in `poly.tolbd` is motivated by the possibility of showing the fit and the tolerance bounds graphically. Such output illustrations are shown in Figure 20 for $k = 1$ and $k = 2$ for some example data taken from Graybill (1976), pp. 274-276.

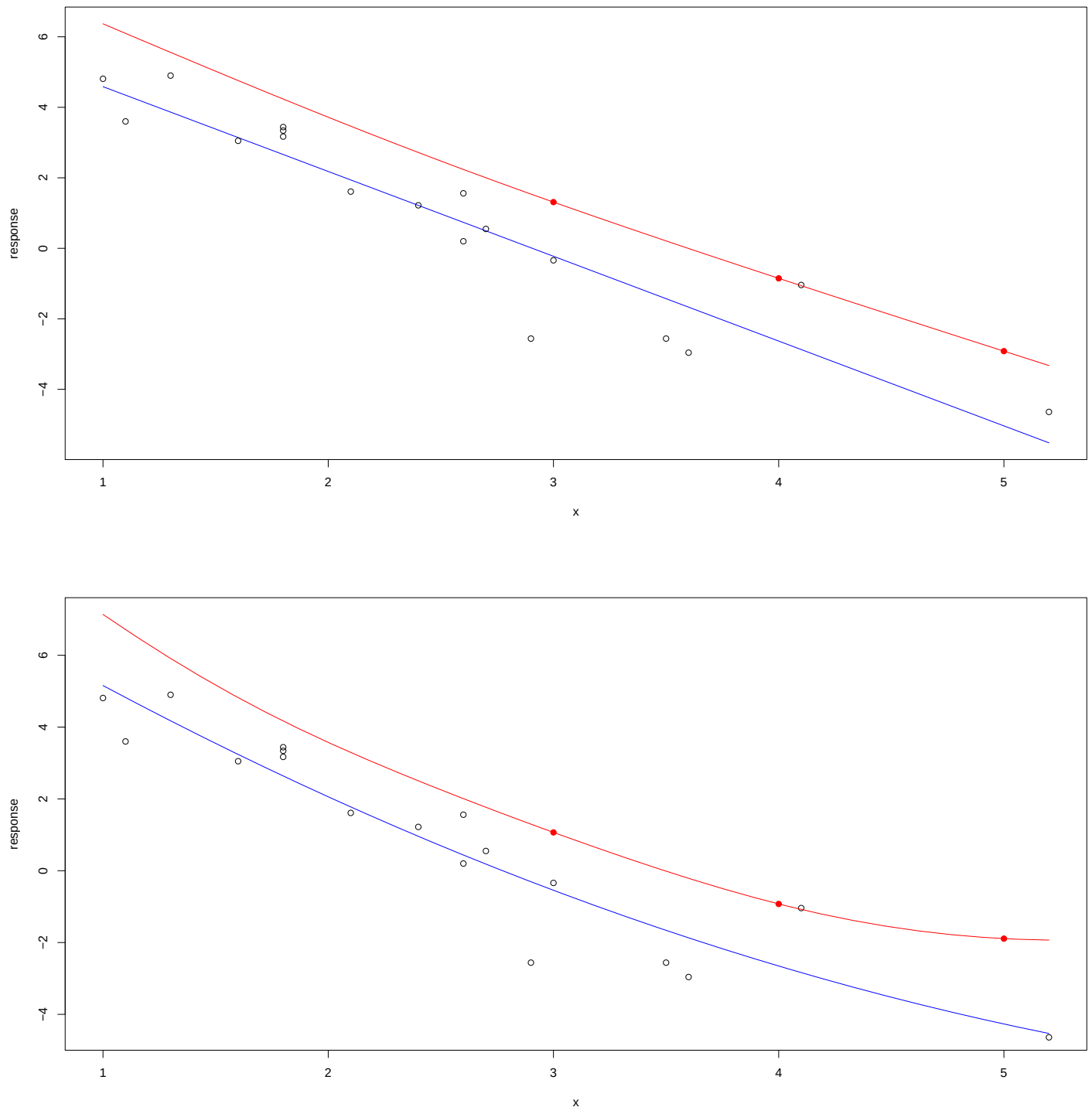






Figure 20: Linear and Quadratic Regression Tolerance Bounds

95% Upper Confidence Bounds for .8-Quantile.

**References.**


1. Amos, D.E. (1965), "Representations of the central and noncentral t distributions," *Biometrika, 51:451–458.*

2. Chou, Y.M., Owen, D.B. and Borrego, S.A. (1990), "Lower Confidence Limits on Process Capability Indices" *Journal of Quality Technology*, 22:223–229.

3. Cooper, B.E. (1968), "Algorithm AS5: The integral of the non–central t–distribution," *Appl. Statist.,* 17:224–228.

4. D'Agostino, R.B. and M.A. Stephens, M.A. (1986), *Goodness-of-Fit Techniques*, Marcel Dekker, New York.

5. Daniel, C. and Wood, F.S. (1971), *Fitting Equations to Data*, Wiley, New York.

6. Graybill, F.A. (1976), *Theory and Application of the Linear Model*, Duxbury Press, North Scituate, Massachusetts.

7. Johnson, N.L. and Kotz, S. (1972), *Continuous Univariate Distributions,* Vol. 2. Wiley, New York.

8. Kotz, S. and Johnson, N.L.(1993), *Process Capability Indices,* Chapman & Hall, London.

9. Kotz, S. and Lovelace, C.R. (1998), *Process Capability Indices in Theory and Practice,* Arnold, London.

10. Lehmann, E.L. and Romano, J.P. (2005), *Testing Statistical Hypotheses, Third Edition*, John Wiley & Sons, New York.

11. Odeh, R.E. and Owen, D.B. (1980), "Tables for normal tolerance limits, sampling plans, and screening," *Marcel Dekker,* New York.

12. Owen, D.B. (1968), "A survey of properties and applications of the noncentral t–distribution, *Technometrics,* 10:445–478.

13. Owen, D.B. (1985), "Noncentral t–distribution," *Encyclopedia of Statistical Sciences,* Vol. 6. Wiley, New York.

14. Scholz, F.W. and Vangel, M. (1998), "Tolerance Bounds and $C_{pk}$ Confidence Bounds Under Batch Effects," in *Advances in Stochastic Models for Reliability, Quality and Safety, eds. W. Kahle, E. von Collani, J. Franz, and U. Jensen*, Birkhäuser, Boston.

## Appendix A: Equivalence of Tail Probability and Quantile Bounds

We will start with a given $x$ which defines $p = p(x) = P(X \leq x)$. The $100\gamma\%$ lower confidence bound for $x_p = x_{p(x)} = x$ is given by $\bar{X} - k_p S$ where $k_p$ solves $G_{n-1,-\sqrt{n}\, z_p}(\sqrt{n}k) = \gamma$ for $k$. Since $-\sqrt{n}z_p$ is strictly decreasing in $p$ we have that $G_{n-1,-\sqrt{n}\, z_p}(y)$ is strictly increasing in $p$ for any fixed $y$ and $k = k_p$ as solution to $G_{n-1,-\sqrt{n}\, z_p}(\sqrt{n}k) = \gamma$ is a strictly decreasing function of $p$. Thus $h(p) = -k_p$ is strictly increasing in $p$. For $p = p(x)$ we have

$$\gamma = P(\bar{X} - k_p S \leq x_p) = P(\bar{X} + h(p)S \leq x) = P\left(h(p) \leq \frac{x - \bar{X}}{S}\right) = P\left(p \leq h^{-1}\left(\frac{x - \bar{X}}{S}\right)\right) .$$

Thus $\tilde{p}_U(x) = h^{-1}((x - \bar{X})/S)$ is a $100\gamma\%$ upper confidence bound for $p = p(x)$. We now show that it coincides with the upper bound $\hat{p}_U = \hat{p}_U(x)$ defined in the main text.

Rather than using the cumbersome $\tilde{p}_U(x)$ as subscript in several of the following steps we write $q = \tilde{p}_U(x)$ for short. This $q$ satisfies $h(q) = (x - \bar{X})/S$. Since $h(q) = -k_q$, with $k = k_q$ being the solution to

$$\gamma = G_{n-1,-\sqrt{n}\, z_q}(\sqrt{n}\, k) = G_{n-1,-\sqrt{n}\, z_q}(-\sqrt{n}\, h(q)) = 1 - G_{n-1,\sqrt{n}\, z_q}(\sqrt{n}\, h(q))$$

where the last equality comes from identity (3), this rewrites as

$$1 - \gamma = G_{n-1,\sqrt{n}\, z_q}(\sqrt{n}\, h(q)) = G_{n-1,\sqrt{n}\, z_q}\left(\sqrt{n}\, \frac{x - \bar{X}}{S}\right)$$

and solving this for $q$ yields $q = \hat{p}_U(x)$ according to the definition of $\hat{p}_U(x)$. Thus $\tilde{p}_U(x) = \hat{p}_U(x)$.

It should now be quite clear from the above gyrations (which may appear to be a slight of hand) why the direct approach was taken in defining $\hat{p}_U(x)$ rather than using the above form $\tilde{p}_U(x)$, which leads to a double root solving process, namely solving $h(p) = (x - \bar{X})/S$ for $p$ and then finding $k_p = -h(p)$ (by root solving from $G_{n-1,-\sqrt{n}\, z_p}(\sqrt{n}k) = \gamma$) for each evaluation of $h(p)$.

## Appendix B: Mean and Variance of $SS_T = SS_b$ when $\sigma_e = 0$

When $\sigma_e = 0$ then $SS_T = SS_b = \sum_i n_i(\bar{X}_{i\cdot} - \bar{X})^2 = N\sum_i w_i(\bar{X}_{i\cdot} - \bar{X})^2$ with $w_i = n_i/N$. Note that $\bar{X}_{i\cdot} = \mu + b_i$ and $\bar{X} = \sum_i w_i \bar{X}_{i\cdot} = \mu + \sum_i w_i b_i$. Then we can write

$$\begin{aligned} SS_b &= N\sum_i w_i(b_i - \sum_j w_j b_j)^2 = N\left[\sum_i w_i b_i^2 - 2\sum_i w_i b_i \sum_j w_j b_j + \sum_i w_i (\sum_j w_j b_j)^2\right] \\ &= N\left[\sum_i w_i b_i^2 - 2(\sum_j w_j b_j)^2 + (\sum_j w_j b_j)^2\right] = N\left[\sum_i w_i b_i^2 - (\sum_j w_j b_j)^2\right] \end{aligned}$$

Since $\sum_j w_j b_j \sim \mathcal{N}(0, \sigma_b^2 \sum_i w_i^2)$ we have

$$E(SS_b) = N\left[\sum_i w_i E(b_i^2) - E\left((\sum_j w_j b_j)^2\right)\right] = N\left[\sum_i w_i \sigma_b^2 - \sigma_b^2 \sum_i w_i^2\right] = N\sigma_b^2\left(1 - \sum_i w_i^2\right) .$$

Next note that $\mathrm{var}(b_i^2) = 2\sigma_b^4$. Exploiting independence of the $b_i$ and $E(b_i) = 0$ we get

$$\mathrm{cov}(b_i^2, (\sum_j w_j b_j)^2) = \sum_j \sum_{j'} w_j w_{j'} \mathrm{cov}(b_i^2, b_j b_{j'}) = w_i^2 \mathrm{var}(b_i^2) = 2w_i^2\sigma_b^4$$

Thus

$$\begin{aligned} \mathrm{var}(SS_b) &= N^2\left[\mathrm{var}\left(\sum_i w_i b_i^2\right) - 2\mathrm{cov}\left(\sum_i w_i b_i^2, (\sum_j w_j b_j)^2\right) + \mathrm{var}\left((\sum_j w_j b_j)^2\right)\right] \\ &= N^2\left[\sum_i w_i^2 \mathrm{var}(b_i^2) - 2\sum_i w_i \mathrm{cov}(b_i^2, (\sum_j w_j b_j)^2) + 2\sigma_b^4 (\sum_i w_i^2)^2\right] \\ &= N^2\left[\sum_i w_i^2 2\sigma_b^4 - 2\sum_i w_i 2w_i^2\sigma_b^4 + 2\sigma_b^4(\sum_i w_i^2)^2\right] = 2N^2\sigma_b^4\left[\sum_i w_i^2 - 2\sum_i w_i^3 + (\sum_i w_i^2)^2\right] \end{aligned}$$

Matching mean and variance of an approximating $a\chi_g^2$ random variable with $E(SS_b)$ and $\mathrm{var}(SS_b)$, respectively gives

$$E(SS_b) = N\sigma_b^2\left(1 - \sum_i w_i^2\right) = E(a\chi_g^2) = ag$$

and

$$\mathrm{var}(SS_b) = 2N^2\sigma_b^4\left[\sum_i w_i^2 - 2\sum_i w_i^3 + (\sum_i w_i^2)^2\right] = \mathrm{var}(a\chi_g^2) = 2a^2 g$$

we get

$$\frac{\mathrm{var}(SS_b)}{[E(SS_b)]^2} = \frac{2N^2\sigma_b^4\left[\sum_i w_i^2 - 2\sum_i w_i^3 + (\sum_i w_i^2)^2\right]}{N^2\sigma_b^4\left(1-\sum_i w_i^2\right)^2} = \frac{2a^2g}{a^2g^2} = \frac{2}{g}$$

$$\Longrightarrow \qquad g = \frac{\left(1-\sum_i w_i^2\right)^2}{\sum_i w_i^2 - 2\sum_i w_i^3 + (\sum_i w_i^2)^2} \qquad \text{and} \qquad a = \frac{N\sigma_b^2(1-\sum_i w_i^2)}{g}$$

for the Satterthwaite approximation $a\chi_g^2$ for $SS_b$.

## Appendix C

Here we present the rationale for the approximation $g \approx f$. Let $w_i = n_i/N$ and observe $\sum_{i=1}^k w_i = 1$. Further let

$$A = \sum_{i=1}^k w_i^2 \quad \text{and} \quad U = \sum_{i=1}^k w_i\,(w_i - A)^2 = \sum_{i=1}^k w_i^3 - A^2\ .$$

Then

$$g = \frac{(1-\sum w_i^2)^2}{\sum w_i^2 - 2\sum w_i^3 + (\sum w_i^2)^2} = \frac{1-A}{A}\,\frac{1-A}{1-A-2U/A} \approx \frac{1-A}{A} = f\ ,$$

where in the approximation step we assume that

$$\frac{U}{A} = \sum_{i=1}^k w_i\left(\frac{w_i}{A} - 1\right)^2 A \ \ll\ 1\ , \quad \text{since} \quad w_i \approx \frac{1}{k}\,, \quad A \approx \frac{1}{k}\,, \quad \frac{w_i}{A} \approx 1\ .$$

Note that $U/A = 0$, when the $n_i$ are all the same. In that case the above approximation is exact.