Power and Sample Size Calculations in JMP®

Clay Barker, Senior Research Statistician Developer

Summary. JMP provides tools for performing prospective power and sample size calculations for a variety of settings. These capabilities range from aiding the user in planning a basic test on the mean or variance, to more complicated situations such as designing a reliability study. This document provides details on the theory behind the power and sample size features in JMP.

1. Introduction

JMP provides tools for conducting prospective power and sample size calculations for a variety of settings. This document provides details on the strategies and formulas that are used in the JMP Power and Sample Size Calculator. Sample JSL functions are also provided to mimic what is happening behind the scenes in JMP. Section 2 provides details on calculations for planning tests on normal means and variances. Section 3 provides details on calculations for planning tests on a binomial proportion. Section 4 covers some basic notation for reliability data that will be useful in Section 5 and Section 6, which cover reliability demonstration plans and reliability test plans respectively. Closing comments are in Section 7, and a convenient summary of formulas appears in Appendix A.

2. Calculations for normal means and variances

The Power and Sample Size Calculator in JMP provides functionality for planning tests on normal means and variances. This section summarizes the strategies employed by JMP for performing such calculations. Topics covered in this section include: test of k sample means, test of a standard deviation, and a test on the number of counts per unit. For each topic, we discuss the motivation behind the methods used by JMP and provide sample JSL code for replicating JMP results.

2.1. One Sample Mean Test

Suppose we have a random sample of n normal variates, $X_i \sim \mathrm{N}(\mu, \sigma^2)$. We want to test the hypothesis

$$\mathrm{H}_0 : \mu = \mu_0 \text{ vs } \mathrm{H}_a : \mu \neq \mu_0.$$

One way to conduct this test is to use the traditional full-vs-reduced F test

$$F^\star = \frac{\text{SSE(Reduced)-SSE(Full)}}{\text{df(Reduced)-df(Full)}} \times \frac{1}{\text{MSE(Full)}} \tag{1}$$

which has an $\mathrm{F}(1, n-1)$ distribution under the null hypothesis. Here the full model uses the sample mean for μ and the reduced model uses μ_0 for μ, leading to

$$\begin{aligned} \text{SSE(Reduced)} &= \sum_{i=1}^{n}(x_i - \mu_0)^2 \\ \text{SSE(Full)} &= \sum_{i=1}^{n}(x_i - \bar{x})^2. \end{aligned}$$

Under the alternative hypothesis, we know that $F^\star$ has an $\mathrm{F}(1, n-1, n\delta^2/\sigma^2)$ distribution where $n\delta^2/\sigma^2$ is a noncentrality parameter. Here δ is an assumed difference between the true mean and the hypothesized mean μ_0.

How do we use $F^\star$ to do power and sample size calculations? To calculate power, first we need to get a critical value for testing H_0. Given a testing level α,

$$f_{\text{crit}} = \Phi^{-1}(1-\alpha, 1, n-1)$$

where $\Phi^{-1}(1-\alpha, 1, n-1)$ is the inverse CDF of the $\mathrm{F}(1, n-1)$ distribution evaluated at $1-\alpha$. So to calculate power, we use

$$\begin{aligned} \text{power} &= \Pr(\text{reject } \mathrm{H}_0 \text{ given } \mu = \mu_0 + \delta) \\ &= \Pr\left[f > f_{\text{crit}} \,\middle|\, f \sim \mathrm{F}\left(1, n-1, \frac{n\delta^2}{\sigma^2}\right)\right] \end{aligned} \tag{2}$$

given sample size n, testing level α, error variance σ^2, and difference to detect δ.

We can use built-in JSL functions to calculate power directly. For $n = 20$, $\delta = .75$, and $\sigma = 1.25$, we use the following example JSL function to calculate power.

```
f_power = function({n, delta, sigma, alpha=.05},{fcrit, dsr, pow},
  fcrit = fquantile(1-alpha, 1, n-1);
  dsr = delta/sigma;
  pow = 1 - f distribution(fcrit, 1, n-1, n*dsr*dsr);
  pow;
);

f_power(20, .75, 1.25);
```

Given (2), obtaining a sample size that yields power of $1-\beta$ is straightforward. We could start at $n = 10$, for example, and increase n until the power is greater than or equal to $1-\beta$. The following JSL function uses the power function that we just defined to calculate sample size.

```
f_ss = function({delta, sigma, its, dpower, alpha}, {n, i, pow},
  n = 10; // place to start
  for(i=1, i<=its, i++,
    pow = f_power(n, delta, sigma, alpha);
    if(pow>=dpower, break());
    n++;
  );
  n;
);

f_ss(.75, 1.25, 1000, .95, .05);
```

Fortunately JMP uses a more efficient strategy based on (2) than this brute force method for determining sample size. JMP does not provide the option to plan a one-sided test of the mean. But we can obtain calculations for the one-sided case simply by doubling the desired α level and using the Power and Sample Size Calculator in JMP. So if we wanted to do a one-sided sample size calculation with $\alpha = .05$, we would use a two-sided calculation with $\alpha = .1$.

2.1.1. Distributional Results

How do we arrive at the distributional results for $F^\star$ stated earlier? First we realize that under the null and alternative hypotheses, SSE(Full)/σ^2 has a χ^2_{n-1} distribution. Then we rewrite the difference in sums of squares:

$$\begin{aligned}
\text{SSE(Reduced)} - \text{SSE(Full)} &= \sum_{i=1}^{n}(x_i - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(x_i - \bar{x})^2 + n(\bar{x} - \mu_0)^2 - \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= n(\bar{x} - \mu_0)^2.
\end{aligned}$$

Under the null hypothesis, $\mu = \mu_0$ and

$$\begin{aligned}
\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} &\sim \text{N}(0,1) \\
\Rightarrow n\left(\frac{\bar{x} - \mu_0}{\sigma}\right)^2 = \left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)^2 &\sim \chi^2_1.
\end{aligned}$$

So under the null hypothesis, $F^\star$ is the ratio of χ^2 variates divided by their degrees of freedom. The resulting distribution is F(1,$n-1$).

Similarly under the alternative hypothesis, $\mu_{\text{true}} = \mu_0 + \delta$ and

$$\begin{aligned}
\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} &\sim \text{N}(\delta,1) \\
\Rightarrow n\left(\frac{\bar{x} - \mu_0}{\sigma}\right)^2 = \left(\frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}\right)^2 &\sim \chi^2_1\left(n\delta^2/\sigma^2\right)
\end{aligned}$$

where $\chi^2_1\left(n\delta^2/\sigma^2\right)$ is a non-central χ^2 distribution with non-centrality parameter $n\delta^2/\sigma^2$. So under the alternative hypothesis, $F^\star$ is the ratio of a non-central χ^2 variate (divided by the appropriate degrees of freedom) and a central χ^2 variate (divided by the appropriate degrees of freedom). The resulting distribution is the non-central F distribution: F(1,$n-1$,$n\delta^2/\sigma^2$).

2.2. Two Samples: Comparing Means

Suppose we have two independent samples (each of size n) with common variance σ^2

$$X_{i1} \sim \text{N}(\mu_1, \sigma^2) \text{ and } X_{i2} \sim \text{N}(\mu_2, \sigma^2)$$

and we are interested in testing that the means are equal:

$$\text{H}_0 : \mu_1 - \mu_2 = 0 \text{ vs } \text{H}_a : \mu_1 - \mu_2 \neq 0.$$

One way to test this hypothesis is to use (1) again. In this case we have

$$\text{SSE(Full)} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 + \sum_{i=1}^{n}(x_{i2} - \bar{x}_2)^2 \text{ where } \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$$

and

$$\text{SSE(Reduced)} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_0)^2 + \sum_{i=1}^{n}(x_{i2} - \bar{x}_0)^2 \text{ where } \bar{x}_0 = \frac{\bar{x}_1}{2} + \frac{\bar{x}_2}{2}.$$

The difference in sums of squares can be written

$$\begin{aligned}\text{SSE(R) - SSE(F)} &= \sum_{i=1}^{n}(x_{i1}-\bar{x}_0)^2+\sum_{i=1}^{n}(x_{i2}-\bar{x}_0)^2-\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2 \\ &\quad -\sum_{i=1}^{n}(x_{i2}-\bar{x}_2)^2 \\ &= n(\bar{x}_1-\bar{x}_0)^2+n(\bar{x}_2-\bar{x}_0)^2 \\ &= \frac{n}{4}(\bar{x}_1-\bar{x}_2)^2+\frac{n}{4}(\bar{x}_2-\bar{x}_1)^2 \\ &= \frac{n}{2}(\bar{x}_1-\bar{x}_2)^2.\end{aligned}$$

Because of independence, we know that $\bar{x}_1-\bar{x}_2 \sim \mathrm{N}(\mu_1-\mu_2, 2\sigma^2/n)$. Then under the null hypothesis $\mu_1=\mu_2$ and we have

$$\begin{aligned}\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{2/n}} &\sim \mathrm{N}(0,1) \\ \Rightarrow \left[\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{2/n}}\right]^2 &\sim \chi_1^2.\end{aligned}$$

Under the alternative hypothesis, $\mu_1-\mu_2=\delta$ and

$$\begin{aligned}\bar{x}_1-\bar{x}_2 &\sim \mathrm{N}(\delta, 2\sigma^2/n) \\ \Rightarrow \left[\frac{\bar{x}_1-\bar{x}_2}{\sigma\sqrt{2/n}}\right]^2 &\sim \chi_1^2\left(\frac{n\delta^2}{2\sigma^2}\right)\end{aligned}$$

where $n\delta^2/(2\sigma^2)$ is a non-centrality parameter.

We know that under the null and alternative hypotheses

$$\frac{\text{SSE(Full)}}{\sigma^2} \sim \chi^2_{2n-2},$$

so $F^\star$ has a central F-distribution under the null hypothesis and a non-central F-distribution under the alternative hypothesis:

$$F^\star \sim \mathrm{F}(1, 2n-2) \text{ under } \mathrm{H}_0$$

$$\text{and } F^\star \sim \mathrm{F}\left(1, 2n-2, \frac{n\delta^2}{2\sigma^2}\right) \text{ under } \mathrm{H}_a.$$

We can use this information to perform power calculations. Given testing level α,

$$f_{\text{crit}} = \Phi^{-1}(1-\alpha, 1, 2n-2)$$

where $\Phi^{-1}(1-\alpha, 1, 2n-2)$ is the inverse CDF of the F(1,2n-2) distribution evaluated at $1-\alpha$. Then the power is expressed

$$\begin{aligned}\text{power} &= \text{Pr(reject H}_0\text{ given } \mu_1-\mu_2=\delta\text{)} \\ &= \text{Pr}\left[f > f_{\text{crit}} \,\middle|\, f \sim \mathrm{F}\left(1, 2n-2, \frac{n\delta^2}{2\sigma^2}\right)\right]\end{aligned}$$

for constant sample size n, testing level α, error variance σ^2, and assumed difference in means δ.

We can write our own JSL function to calculate power to detect a difference in means. For two independent samples of size $n = 25$ each, with error variance $\sigma^2 = 1.25^2$, and assumed difference in means $\delta = .8$, the following JSL function tells us that we have about a 60% chance of rejecting the null hypothesis.

```
f_power = function({n, delta, sigma, alpha=.05},{fcrit, dsr, pow},
  fcrit = f quantile(1-alpha, 1, 2*n-2);
  dsr = delta/sigma;
  pow = 1 - f distribution(fcrit, 1, 2*n-2, (n/2)*dsr*dsr);
  pow;
);

f_power(25, .8, 1.25);
```

This power function can then be used to solve for the sample size required to achieve a desired power level.

2.3. k Samples: Testing for a Common Mean

If we have k independent samples (each of size n) of random normals

$$X_{ij} \sim \mathrm{N}(\mu_j, \sigma^2) \text{ for } i = 1, \ldots, n; \; j = 1, \ldots, k$$

we may be interested in testing whether or not the means are all equal

$$\mathrm{H}_0 : \mu_1 = \mu_2 = \ldots = \mu_k \text{ vs } \mathrm{H}_a : \text{ not all means equal.}$$

Once again, we can test this hypothesis using (1). Using techniques similar to the previous two subsections, we can show that

$$\begin{aligned} F^\star &\sim \mathrm{F}\left(k-1, kn-k\right) \text{ under } \mathrm{H}_0 \\ \text{and } F^\star &\sim \mathrm{F}\left(k-1, kn-k, \frac{n\sum_{j=1}^{k}(\mu_j - \bar{\mu})^2}{\sigma^2}\right) \text{ under } \mathrm{H}_a, \end{aligned}$$

where $\bar{\mu} = \sum_{j=1}^{k} \mu_j / k$.

For example, suppose that we have four samples ($n = 20$ each) of random normals and want to test to see if all of the means are equal. We will assume that $\sigma^2 = 4^2 = 16$ and we are interested in the power to detect a difference in means when $\mu_1 = 10$, $\mu_2 = 11$, $\mu_3 = 13$, and $\mu_4 = 14$. We can use the following JSL function to find that we will have about an 84% chance of concluding that the means are not all equal.

```
k_power = function({nper, muvec, sigma, alpha=.05},
                  {k, fcrit, mubar, sigma2, del},
  k = nrow(muvec);
  fcrit = f quantile(1-alpha, k-1, nper*k-k);
  mubar = mean(muvec);
  sigma2 = sigma*sigma;
  del = sum( (muvec-mubar)`*(muvec-mubar) )*nper/sigma2;
  1-f distribution(fcrit, k-1, nper*k-k, del);
);

k_power(20, [10, 11, 13, 14], 4);
```

This power function could then be used to determine the sample size needed to obtain a desired power level.

2.4. One Sample Test of Standard Deviation

Suppose we have a sample of n random normal variates with mean μ and variance σ^2: $X_i \sim \mathrm{N}(\mu, \sigma^2)$. We are interested in testing the hypothesis

$$\mathrm{H}_0 : \sigma = \sigma_0 \text{ vs } \mathrm{H}_a : \sigma > \sigma_0 \tag{3}$$

using the test statistic

$$T = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2}.$$

Large values of T provide evidence against the null hypothesis. Under the null hypothesis, $\sigma = \sigma_0$ and $T \sim \chi^2_{n-1}$. Under the alternative hypothesis, the true standard deviation is greater than σ_0 and

$$\begin{aligned} T = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma_0^2} &= \underbrace{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\sigma^2}}_{\sim\chi^2_{n-1}} \times \frac{\sigma^2}{\sigma_0^2} \\ \Rightarrow \frac{\sigma_0^2}{\sigma^2} T &\sim \chi^2_{n-1} \\ \Rightarrow \frac{\sigma_0^2}{(\sigma_0 + \delta)^2} T &\sim \chi^2_{n-1} \end{aligned}$$

where $\delta = \sigma - \sigma_0$. So under the alternative hypothesis, T does not have a χ^2 distribution but is instead a constant multiple of a χ^2 distribution. Then the power to detect this difference in standard deviation is written

$$\begin{aligned} \text{power} &= \text{Pr(reject H}_0 \text{ given } \sigma = \sigma_0 + \delta) \\ &= \Pr\left(X > \frac{\sigma_0^2}{(\sigma_0 + \delta)^2} \times \chi_{1-\alpha}\right) \end{aligned}$$

where $\chi_{1-\alpha}$ is the $1 - \alpha$ quantile of the χ^2_{n-1} distribution and $X \sim \chi^2_{n-1}$.

For example, suppose we want to test (3) at the $\alpha = .05$ level with $n = 20$ observations when $\sigma_0 = 1.2$ and we want to detect a difference of at least .1 (when the true standard deviation is at least 1.3). The following JSL function tells us that we only have about a 14% chance of rejecting the null hypothesis.

```
sd_power_larger = function({sig0, delta, n, alpha=.05}, {crit, spd},
  crit = chi square quantile(1-alpha, n-1);
  spd = sig0+delta;
  1-chi square distribution(crit*sig0*sig0/(spd*spd), n-1);
);

sd_power_larger(1.2, .1, 20);
```

So we should consider increasing the sample size in this example in order to have greater power to reject the null hypothesis. Calling this JSL function repeatedly, we could increase n until we reach a more acceptable power level.

We can also rearrange the expression for power to solve for δ. Solving for δ tells us the difference that we would be able to detect given a specified sample size and power level. Using $1-\beta$ to denote power,

$$\begin{aligned}
1-\beta &= \Pr\left(X > \frac{\sigma_0^2}{(\sigma_0+\delta)^2} \times \chi_{1-\alpha}\right) \text{ where } X \sim \chi^2_{n-1} \\
\Rightarrow \chi_\beta &= \frac{\sigma_0^2}{(\sigma_0+\delta)^2}\chi_{1-\alpha} \\
\Rightarrow \delta &= \sigma_0\sqrt{\frac{\chi_{1-\alpha}}{\chi_\beta}} - \sigma_0 \qquad (4)
\end{aligned}$$

where χ_p is the p^{th} quantile of the χ^2_{n-1} distribution. For example, we may be interested in the magnitude of the change from $\sigma_0 = 2$ that we can detect with 90% power when we have $n = 50$ observations. The following JSL function tells us that we could detect an increase of about $\delta = .68$.

```
delta_larger = function({sigma0, pow, n, alpha=.05},{chi1, chi2},
  chi1 = chi square quantile(1-alpha, n-1);
  chi2 = chi square quantile(1-pow, n-1);
  sigma0*sqrt(chi1/chi2) - sigma0;
);

delta_larger(2, .9, 50);
```

Instead of (3), we may be interested in testing

$$\text{H}_0 : \sigma \geq \sigma_0 \text{ vs } \text{H}_a : \sigma < \sigma_0 \qquad (5)$$

using the same test statistic T that we used when testing (3). Now small values of T provide evidence against the null hypothesis. The power to reject the null hypothesis is

$$\begin{aligned}
\text{power} &= \Pr(\text{reject H}_0 \text{ given } \sigma = \sigma_0 - \delta) \\
&= \Pr\left(X < \frac{\sigma_0^2}{(\sigma_0-\delta)^2} \times \chi_\alpha\right)
\end{aligned}$$

where χ_α is the α quantile of the χ^2_{n-1} distribution, $\sigma + \delta = \sigma_0$, and $X \sim \chi^2_{n-1}$.

For example, suppose we want to test (5) at the $\alpha = .05$ level with $n = 20$ observations when $\sigma_0 = 2$ and we want to detect a difference of $\delta = .5$ (so true standard deviation of 1.5). The following JSL function tells us that the power to reject the null hypothesis is about 48%.

```
sd_power_smaller = function({sig0, delta, n, alpha=.05},{crit, smd},
  crit = chi square quantile(alpha, n-1);
  smd = sig0-delta;
  chi square distribution(crit*sig0*sig0/(smd*smd), n-1 );
);

sd_power_smaller(2, .5, 20);
```

Again, we could increment n and use this JSL function to help us find the sample size needed to have a desired power to reject the null hypothesis.

Similar to (4), we are able to use

$$\delta = \sigma_0 \sqrt{\frac{\chi_\alpha}{\chi_{1-\beta}}} - \sigma_0$$

to solve for δ when given n, σ_0, and a desired power.

2.5. Counts per Unit

Suppose we are interested in the number of defects (or any other meaningful count) per unit of some product. For example, we may be interested in the number of leather imperfections on each basketball that we produce. Since we are working with counts, we can assume that the number of defects for unit i, d_i, has a Poisson distribution with parameter λ

$$d_i \sim \text{Poisson}(\lambda) \Rightarrow \text{E}(d_i) = \text{Var}(d_i) = \lambda.$$

We are interested in detecting a change in the average number of defects per unit. For convenience we will focus on detecting an increase in the average, which means testing the hypothesis

$$\text{H}_0 : \lambda \leq \lambda_0 \text{ vs } \text{H}_a : \lambda > \lambda_0. \tag{6}$$

One way to test (6) would be to use the test statistic

$$T = \frac{\bar{d} - \lambda_0}{\sqrt{\lambda_0/n}}$$

where n is our sample size and $\bar{d}$ is the mean number of defects per unit in our sample. Because of the central limit theorem, we assume that T is approximately standard normal under the null hypothesis.

To calculate power to detect an increase in the mean counts per unit, we need to find the distribution of T under the alternative hypothesis. Noting that $\lambda = \lambda_0 + \delta$ under the alternative, we can rearrange the test statistic as

$$\begin{aligned}
T &= \frac{\bar{d}}{\sqrt{\lambda_0/n}} - \sqrt{n\lambda_0} \\
&= \frac{\bar{d}}{\sqrt{\lambda/n}} \times \frac{\sqrt{\lambda/n}}{\sqrt{(\lambda-\delta)/n}} - \sqrt{n\lambda_0} \\
&= \underbrace{\frac{\bar{d}-\lambda}{\sqrt{\lambda/n}}}_{\sim N(0,1)} \times \sqrt{\frac{\lambda}{\lambda-\delta}} + \delta\sqrt{\frac{n}{\lambda-\delta}}.
\end{aligned}$$

Rewriting T in this manner reveals that

$$T \sim \text{N}\left(\delta\sqrt{\frac{n}{\lambda_0}}, \frac{\lambda_0+\delta}{\lambda_0}\right)$$

under the alternative hypothesis of the true mean being $\lambda_0 + \delta$. Now we can use this result to perform power and sample size calculations. If we are performing an α level test, the power to reject the null hypothesis is written

$$\begin{aligned} \text{power} &= \text{Pr(reject H}_0 \text{ given } \lambda = \lambda_0 + \delta) \\ &= \Pr\left(T > Z_{1-\alpha} \,\middle|\, \lambda = \lambda_0 + \delta\right) \\ &= 1 - \Phi\left(\frac{Z_{1-\alpha} - \delta\sqrt{n/\lambda_0}}{\sqrt{(\lambda_0+\delta)/\lambda_0}}\right) \qquad (7) \end{aligned}$$

where $Z_{1-\alpha}$ is the $1-\alpha$ quantile of the standard normal distribution and $\Phi()$ is the cumulative distribution function of the standard normal distribution.

For example, say that we are interested in testing that the mean number of defects per unit is $\lambda_0 = 4$ and we want to know the power to detect an increase of one defect per unit. We are limited to a sample of $n = 50$ units. We can use the following JSL function to determine that we have about a 55% chance of detecting this increase.

```
cpu_power = function({lambda0, del, n, alpha=.05},{z, pow},
  z = normal quantile(1-alpha);
  pow = 1 - normal distribution( (z -
    del*sqrt(n/lambda0))/sqrt((lambda0+del)/lambda0) );
  pow;
);

l_0 = 4;
delta = .5;
n=50;
cpu_power(l_0, delta, n);
```

We can also use (7) to solve for the sample size necessary to obtain a specified power to reject the null hypothesis. Using $1 - \beta$ to denote the desired power to reject the null hypothesis, the sample size is

$$n = \frac{\lambda_0}{\delta^2}\left(Z_{1-\alpha} - Z_\beta\sqrt{\frac{\lambda_0+\delta}{\lambda_0}}\right)^2$$

where Z_{1-p} is the $1 - p$ quantile of the standard normal distribution.

For example, suppose we want to test that the average number of defects per unit is $\lambda_0 = 4$ and we want to find the sample size needed to provide 90% power to detect an increase of one defect per unit. Using the following JSL function, we find that we need $n = 38$ units to have the specified power to detect the increase.

```
cpu_n = function({lambda0, del, pow, alpha=.05},{za, zb, const, n},
  za = normal quantile(1-alpha);
  zb = normal quantile(1-pow);
  const = (za-zb*sqrt((lambda0+del)/lambda0))/del;
  n = lambda0*const*const;
  ceiling(n);
);

l_0 = 4;
delta = 1;
pow = .9;
cpu_n(l_0, delta, pow);
```

3. Calculations for a binomial proportion

When conducting an experiment where the response is binomially distributed, researchers are typically interested in testing the hypothesis

$$\mathrm{H}_0 : p = p_0 \text{ vs } \mathrm{H}_a : p \neq p_0 \tag{8}$$

or a one-sided version of the alternative. When planning such a study, the normal approximation to the binomial distribution provides a simple way to calculate the sample size, n, necessary to have power $1-\beta$ to reject the null hypothesis. For small sample sizes, or more specifically when np or $n(1-p)$ is small, the normal distribution cannot adequately approximate the binomial distribution. This results in unwarranted optimism. To avoid these mistaken conclusions, we advocate the use of exact power calculations. While exact power calculations do not have as simple a form as the normal approximation, they can be calculated efficiently with current computing resources.

3.1. An adjusted test of the binomial proportion

The usual Wald test statistic for the hypothesis stated in (8) is

$$T(y) = \frac{(\hat{p}-p_0)^2}{\frac{\hat{p}(1-\hat{p})}{n}} \text{ where } \hat{p} = y/n. \tag{9}$$

Agresti and Coull (1998) discuss some of the weaknesses of Wald-based tests and confidence intervals for p. Wald tests of p are often liberal, meaning that the Type I error rate is higher than the nominal level.

Because of the poor performance of the Wald-based tests and intervals, Agresti and Coull (1998) suggest using a slight adjustment to the Wald test statistic

$$T(y) = \frac{(\hat{p}-p_0)^2}{\frac{\hat{p}(1-\hat{p})}{n+4}} \text{ where } \hat{p} = \frac{y+2}{n+4}. \tag{10}$$

This simple modification of (9) is motivated by the score test for p, but is more simple to use than the score statistic. The adjusted Wald test statistic performs better than the Wald test statistic here in terms of maintaining the desired Type I error rate for tests and coverage rates for confidence intervals. Figure 1 shows how Type I error rate varies as a function of p for the Wald and adjusted Wald tests when $n = 25$. The tests were conducted at the 0.05 level and we can see that the Wald test is liberal for nearly all values of p. For p close to zero or one, the Type I error rate can be higher than 50%. The adjusted Wald test comes much closer to achieving the nominal testing level.

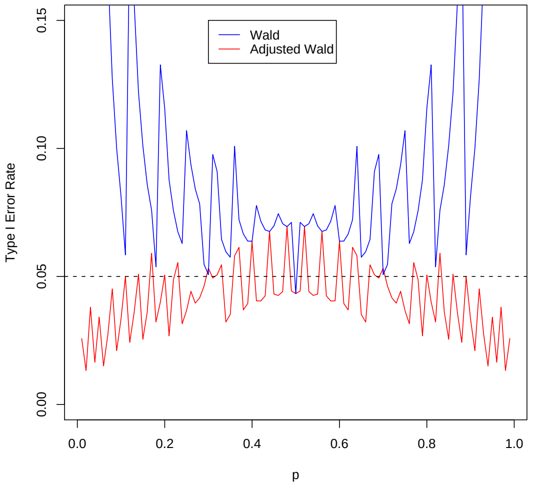

Fig. 1. Plot of Type I error rate as a function of p for $n = 25$.

We consider a simple example to investigate the difference between (9) and (10). Suppose that $Y \sim \text{Binomial}(n = 5, p = 0.005)$ and we want to test the hypothesis $\text{H}_0 : p = 0.1$ vs $\text{H}_a : p \neq 0.1$. If we use (9) to test this hypothesis, then the test statistic goes to infinity when $y = 0$ or $y = n = 5$. So when $y = 0$ is observed, should the null hypothesis be rejected? Intuitively it makes sense to reject since the test statistic is approaching infinity. But under H_0 $(p = 0.1)$, there is a 59% chance of observing $y = 0$. We certainly would not want to use a test with a Type I error rate that high.

Now consider using (10) to test the very same hypothesis. The adjusted Wald test statistic has the favorable quality of taking finite values for y between zero and n. In our simple example, (10) gives us $T(0) = 0.78$ which would not reject at any reasonable level. This example shows that the adjusted test resolves any ambiguity when observations fall at the boundary ($y = 0$ or $y = n$).

3.2. Review of the normal approximation

Here we show how to use the normal approximation to the binomial distribution to determine the appropriate sample size for a test to have a desired power.

3.2.1. Obtaining the approximate sample size

Suppose we want to test the hypothesis stated in (8) at level α. In the previous subsection we discussed the advantages of using the adjusted Wald statistic for testing this hypothesis, but for the sake of simplicity we are going to use the usual Wald statistic for calculating sample size. If we want power $1 - \beta$ to reject the null hypothesis, we can calculate the corresponding sample size as follows

$$\begin{aligned}
\Pr\left(\frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} > \Phi^{-1}(1-\alpha/2)\right) &= 1-\beta \qquad (11)\\
\Pr\left(\frac{\hat{p} - p}{\sqrt{p_0(1-p_0)/n}} + \frac{p - p_0}{\sqrt{p_0(1-p_0)/n}} > \Phi^{-1}(1-\alpha/2)\right) &= 1-\beta\\
\Pr\left(\frac{\hat{p} - p}{\sqrt{p(1-p)/n}} > \Phi^{-1}(1-\alpha/2)\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} - \frac{p - p_0}{\sqrt{p(1-p)/n}}\right) &= 1-\beta
\end{aligned}$$

$$
\begin{aligned}
\Pr\left(z > \Phi^{-1}(1-\alpha/2)\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} - \frac{p-p_0}{\sqrt{p(1-p)/n}}\right) &= 1-\beta \\
\frac{p-p_0}{\sqrt{p(1-p)/n}} - \Phi^{-1}(1-\alpha/2)\sqrt{\frac{p_0(1-p_0)}{p(1-p)}} &= \Phi^{-1}(1-\beta).
\end{aligned}
$$

where $\Phi()$ is the standard normal CDF. Solving for n yields

$$
n = \left[\frac{\Phi^{-1}(1-\beta)\sqrt{p(1-p)} + \Phi^{-1}(1-\alpha/2)\sqrt{p_0(1-p_0)}}{p-p_0}\right]^2. \tag{12}
$$

The National Institute of Standards and Technology (2009) provides the formula in (12). Given the true proportion p, the hypothesized proportion p_0, the testing level α, and desired power, the approximate sample size is a simple function of the normal cumulative distribution function. This formula is appealing because it is simple enough to calculate using a spreadsheet application like Excel. The following JSL function calculates the approximate sample size as a function of p, p_0, power (denoted 'pow' in the JSL function), and testing level α.

```
n_approx = function({p,p0,pow,alpha=.05}, {n},
  n = (sqrt(p*(1-p))*normal quantile(pow) +
   sqrt(p0*(1-p0))*normal quantile(1-alpha/2) )/abs(p-p0);
  n*n;
);
```

A continuity corrected version of (12) also exists for calculating approximate sample size. The derivation of the continuity corrected sample size is nearly identical to that of (12), except the sample proportion is estimated using

$$
\hat{p} = \frac{y+\frac{1}{2}}{n}
$$

and we proceed just as in (11). This modification leads to the approximate sample size

$$
n = \left[\frac{\Phi^{-1}(1-\beta)\sqrt{p(1-p)} + \Phi^{-1}(1-\alpha/2)\sqrt{p_0(1-p_0)}}{p-p_0}\right]^2 + \frac{1}{|p-p_0|}. \tag{13}
$$

Here (13) is just as simple to use as (12), but less prone to underestimating the sample size. The following JSL function uses the function for the raw normal approximation to calculate the continuity corrected sample size.

```
ncc = function({p,p0,pow,alpha=.05}, {},
  n_approx(p,p0,alpha, pow) + 1/abs(p-p0);
);
```

Using the two JSL functions we just defined, we see that the continuity corrected approximation can be quite a bit different from the raw normal approximation. For example, try $p = .995$, $p_0 = .95$, and request 40% power to reject the null hypothesis.

```
show(n_approx(.995, .95, .4));
show(ncc(.995,.95,.4));
```

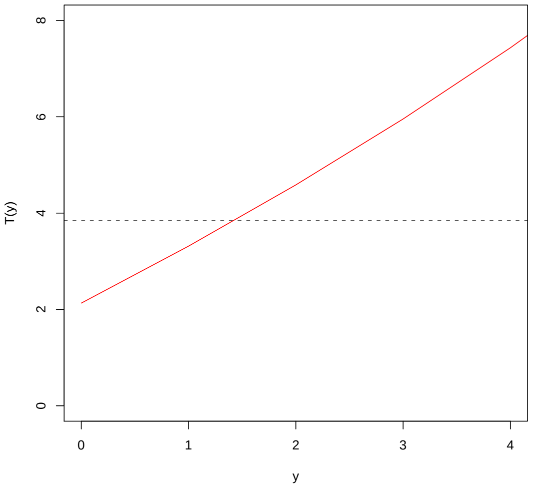

Fig. 2. Plot of adjusted Wald test statistic as a function of y.

3.2.2. A cautionary example

We have seen that both (12) and (13) provide a simple solution for approximate binomial sample sizes. Unfortunately the normal approximation to the binomial breaks down for small n or p. When this happens, the result can be misleading tests (rejecting when you should not) or insufficient sample sizes. The remainder of this section will look at examples of the latter.

Suppose that we want to test $\mathrm{H}_0 : p = 0.0001$ vs $\mathrm{H}_a : p \neq 0.0001$, assuming that the truth is $p = 0.1$. Plugging these values into (12), we find that $n = 27$ is the approximate sample size that gives us 95% power to reject the null hypothesis. The adjusted Wald statistic for testing this hypothesis would look like

$$T(y) = \frac{(\frac{y+2}{31} - 0.0001)^2}{\frac{\frac{y+2}{31}(1-\frac{y+2}{31})}{31}}$$

which has a χ^2_1 distribution under the null hypothesis. Using χ to denote the 1-α quantile of the χ^2_1 distribution, we can solve for the values of Y where the test statistic equals the critical value.

$$T(y) = \chi$$

$$\left(\frac{y+2}{31} - 0.0001\right)^2 = \chi\frac{y+2}{31}\left(1 - \frac{y+2}{31}\right)\frac{1}{31}$$

$$\left(\frac{y+2}{31}\right)^2 - 0.0002\left(\frac{y+2}{31}\right) + 0.0001^2 - \chi\left(\frac{y+2}{31}\right) + \chi\left(\frac{y+2}{31}\right)^2 = 0.$$

Continuing in this line and using the quadratic formula, we find that $T(y) = \chi$ at $y_1 = -2.00$ and $y_2 = 1.42$. These calculations are reflected in Figure 2. This result tells us that the test will reject when $Y \leq y_1$ and when $Y \geq y_2$. Since we are dealing with a binomially distributed random variable, obviously $Y \leq -2.00$ is not possible. So the test only rejects when $Y \geq 1.42$. Since we assumed that the true distribution is binomial(n=27, $p = 0.1$), we can calculate the probability that the test will reject.

$$\begin{aligned} \Pr\{Y \geq 1.42 | Y \sim \text{binomial}(27, 0.1)\} &= \Pr\{Y \geq 2 | Y \sim \text{binomial}(27, 0.1)\} \\ &= 1 - \Pr\{Y \leq 1 | Y \sim \text{binomial}(27, 0.1)\} \\ &\approx 0.767. \end{aligned}$$

Even though we used a nominal 95% power to obtain an approximate sample size, we ended up with a sample size that only provides 76.7% power to reject the null hypothesis. Despite the more precise adjusted Wald test used here, $n = 27$ does not come close to providing the desired power to reject the null hypothesis. Using the continuity corrected estimate instead results in $n = 37$, yielding 89.6% power to reject the null hypothesis. So for this example, (12) is extremely misleading while (13) yields a sample size that comes closer to the nominal power level.

Now suppose that we would like to test $\mathrm{H}_0 : p = 0.1$ vs $\mathrm{H}_a : p \neq 0.1$, with 95% power and assuming that the truth is $p = 0.001$. Formula (12) yields $n = 42$ and formula (13) yields $n = 52$, resulting in approximately 0.0% power and 94.9% power respectively. Here the continuity corrected approximation is able to obtain the desired power level, but (12) yields practically zero power to reject the null hypothesis.

The examples in this section have shown us the potential danger of using the normal approximation to calculate sample sizes needed for testing the binomial proportion. Neither approximation is able to guarantee that the sample size is sufficient to obtain the desired power level. Underestimating the sample size can result in a significant decrease in power. In practice, this could result in erroneous decisions as well as wasted money and resources.

3.3. *Exact power calculations*

We have seen that the normal approximation does not always perform well when used to calculate sample sizes needed for testing the binomial proportion. Instead, exact power calculations should be used to calculate sample size. Exact calculations guarantee that the desired power level is obtained. Here we will look at how to calculate exact power and how to use that information to calculate sample size.

3.3.1. *How to calculate power exactly*

Because the binomial is a discrete distribution, we have the luxury of being able to calculate power exactly using a simple summation. The basic idea is to find all of the values of y that lead to a rejected test. Then add up the probabilities of observing those values of y. More explicitly, let's say that we want to calculate the power to reject the null hypothesis $\mathrm{H}_0 : p = p_0$ when testing at the α level. We will use the adjusted Wald statistic (10) and let χ denote the $1 - \alpha$ quantile of the χ^2_1 distribution. To calculate power, use the summation

$$\text{Power } = \sum_{y=0}^{n} \Pr\{Y = y \mid Y \sim \text{Binomial}(n, p)\} I\{T(y) \geq \chi\} \tag{14}$$

where $I\{T(y) \geq \chi\}$ equals one when $T(y) \geq \chi$ and zero otherwise.

Let's go through a simple example. Suppose we have a sample of $n = 5$ observations and want to test the hypothesis $\mathrm{H}_0 : p = 0.8$ vs $\mathrm{H}_a : p \neq 0.8$. Testing at the $\alpha = 0.05$ level, we will calculate the power to reject the null hypothesis when the true proportion is $p = .1$. Plugging these values into (14) gives us

$$\text{Power} = \sum_{y=0}^{5} \Pr\{Y = y \mid Y \sim \text{Binomial}(5, 0.1)\} I\{T(y) \geq 3.84\}.$$

The calculations are summarized in Table 1 and result in power approximately equal to 0.99. The following JSL functions can be used to perform exact power calculations and double-check the calculations in Table 1.

```
Tstat = function({y,n,p_0}, {phat, T},
  phat = (y+2)/(n+4);
  T = (phat - p_0)/sqrt(phat*(1-phat)/(n+4));
  T;
```

Table 1. Example power calculation.

y	$T(y)$	$I\{T(y) \geq \chi\}$	Pr$(Y = y)$
0	17.4	1	0.59
1	8.8	1	0.33
2	4.6	1	0.07
3	2.2	0	0.008
4	.72	0	4.5e-4
5	.03	0	1e-5

```
);

power_2sided = function({n, p, p_0, level=.05},
                                                {pow, chi, T, Tsqr},
  pow = 0;
  chi = chi square quantile(1-level, 1);
  for(y=0, y<=n, y++,
    T = Tstat(y,n,p_0);
    Tsqr = T*T;
    if( Tsqr > chi, pow += binomial probability(p,n,Y));
  );
  pow;
);
power_2sided(5,.1,.8);
```

3.3.2. *Why did the power drop?*

When working with the binomial distribution, power calculations for testing the proportion behave in a somewhat counter-intuitive manner. We are accustomed to seeing the power to reject the null hypothesis increase monotonically as sample size increases, but this is not the case when working with the binomial distribution. For a particular choice of n, increasing the sample size to $n + 1$ may actually result in a decrease in power. For example, consider testing the hypothesis

$$\mathrm{H}_0 : p = 0.8 \text{ vs } \mathrm{H}_a : p \neq 0.8$$

when the truth is $p = 0.5$. A sample size of eight yields about 64% power to reject the null hypothesis, while a sample size of nine only has 50% power to reject the null hypothesis (tests conducted at the 0.05 level).

Why does this happen? This strange behavior is due to the discreteness of the binomial distribution. For any sample size, we can solve for the values of Y that will lead to rejecting the null hypothesis. In the problem described above, the test rejects for $Y \leq 4.21$ when $n = 8$ and $Y \leq 4.87$ for $n = 9$. Because the binomial is a discrete distribution, this means the test would reject for $Y \leq 4$ for both $n = 8$ and $n = 9$. However,

$$\text{Pr}\{Y \leq 4|Y \sim \text{binomial}(8, 0.5)\} > \text{Pr}\{Y \leq 4|Y \sim \text{binomial}(9, 0.5)\}$$

which means the power drops when going from $n = 8$ to $n = 9$.

Although the behavior may seem quite counter-intuitive, it is not a cause for concern. It does mean that after planning an experiment, researchers should be careful when adding extra runs to the experiment. One option is to add enough runs to the experiment to ensure that the power does not go down; this could mean that a set of two, three, or more runs would be added. Another option would be to not worry about

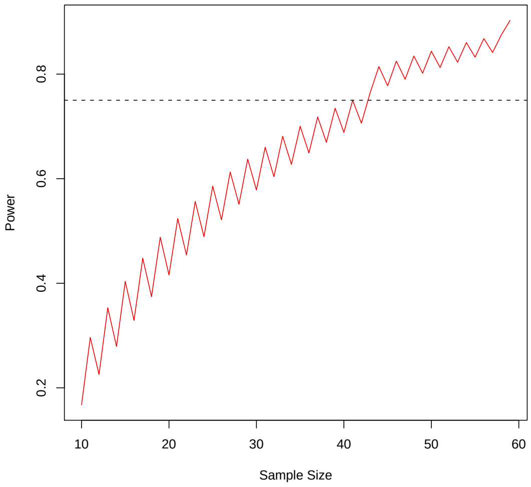

Fig. 3. Plot of power as a function of sample size.

adding a single extra run, since the drop in power is rarely severe. As long as an experiment is carefully planned, the non-monotonic nature of power should not be a cause for concern.

3.3.3. Sample size calculations

In Section 3.2, we looked at examples where the normal approximation did a poor job of calculating sample size. Since it is not hard to calculate power exactly, using exact power calculations to obtain sample size is not a problem. Consider the following situation.

Suppose we have a coin and want to test that the probability of landing on heads is 0.6. That is, we want to test $\mathrm{H}_0 : p = 0.6$ vs $\mathrm{H}_a : p \neq 0.6$. If the true probability is $p = 0.4$, we want 75% power to reject the null hypothesis. So what sample size is needed to obtain this power level? A simple approach would be to start at a particular sample size (say, $n = 10$) and calculate power for a range of values. Once the power gets above 75%, we have an adequate sample size. A more efficient way to do this is to use a bisection search. Figure 3 shows the power as a function of sample size for our example. Judging by the plot, we see that $n = 43$ is adequate to provide 75% power to reject the null hypothesis. So if the true probability of landing on heads is 0.4, we need 43 trials to have at least 75% power to reject the hypothesis $\mathrm{H}_0 : p = 0.6$.

Similarly, given a sample size and desired power, we can calculate the null proportion (p_0) that yields the desired power level. The approach is very similar to calculating sample size.

3.4. Actual testing level

In Section 3.3.3, we saw that in general when solving for sample size, we will not be able to find an n that will yield the desired power precisely. Instead, we find a sample size that yields at least the desired power. Similarly, the actual testing level will be different from the nominal testing level. For example, in Section 3.3.1, we went through an example of how to calculate power for a test of $\mathrm{H}_0 : p = 0.8$ at the $\alpha = .05$ level when $n = 5$. We can go through similar steps to calculate the actual level of this test, this time using 0.8 as the true proportion:

$$\text{Actual } \alpha = \sum_{y=0}^{5} \Pr\{Y = y \mid Y \sim \text{Binomial}(5, 0.8)\} I\{T(y) \geq 3.84\}.$$

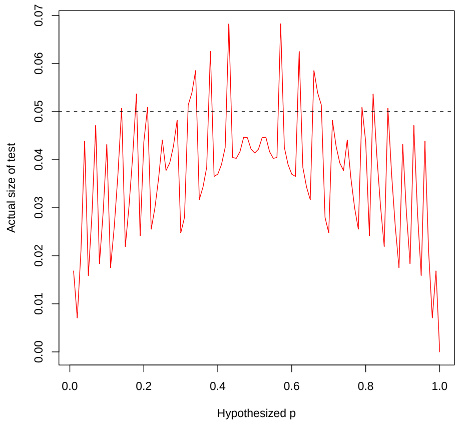

Fig. 4. Plot of actual test size as a function of p_0 for $n = 20$ and nominal testing level $\alpha = .05$.

For this example the actual testing level is 0.0579, slightly higher than the nominal level of 0.05.

Figure 4 again shows that the actual Type I error rate fluctuates around the nominal test size. The actual test size is an important value to consider. If it is too far from the nominal α, one may consider rethinking their experimental set-up. For example if we perform a sample size calculation and find that the actual test size is quite a bit larger than the nominal α, we may consider increasing the sample size until we find an n that achieves the desired power level and yields a type I error rate below the desired level.

3.5. Summary and extension to two-sample proportions

We have looked into issues related to performing sample size calculations for tests of the binomial proportion. Calculations based on the normal distribution are attractive because of their simple form. However, we saw that these approximate values can be misleading, especially when np or $n(1-p)$ is small. A more reliable alternative is to use exact power calculations, which guarantee that the observed power will be greater than or equal to the desired power. Although the approximate sample sizes may perform well in some situations (especially the continuity corrected version), exact calculations are the only way to be sure that the desired power level is obtained. Exact calculations are not quite as convenient as the normal approximation, but they are by no means complicated. In this paper we have only considered the one sample test of proportion, but the same strategies can be easily extended to the two-sample situation testing $\mathrm{H}_0 : p_1 - p_2 = \delta$ vs $\mathrm{H}_a : p_1 - p_2 \neq \delta$. Agresti and Caffo (2000) discuss an adjusted Wald test of the two-sample hypothesis test that is very similar in nature to the test statistic in (10).

4. Reliability Calculations

JMP 9 introduced two new tools for designing reliability studies. Reliability demonstrations allow the reliability of a new product to be compared to a standard. Reliability tests allow us to make inference about either failure probabilities or quantiles of the assumed failure time distribution. This section introduces some basic reliability concepts that will be used in the following sections that provide details about how these features are implemented in the JMP power and sample size platform.

Table 2. Relationships between distributions.

Distribution of x	Cumulative Distribution Function $\Phi(x)$	Distribution of log(x)
Frechet	$\exp\left[-\exp\left(-\frac{\log(x)-\mu}{\sigma}\right)\right]$	Largest Extreme Value
Loglogistic	$\frac{1}{2}+\frac{1}{2}\tanh\left[\frac{\log(x)-\mu}{2\sigma}\right]$	Logistic
Lognormal	$\Phi_N\left[\frac{\log(x)-\mu}{\sigma}\right]$	Normal
Weibull	$1-\exp\left\{-\exp\left[\frac{\log(x)-\mu}{\sigma}\right]\right\}$	Smallest Extreme Value

4.1. Location scale vs. log-location scale

JMP allows the user to assume that failure times take several different forms, either from a location scale distribution or a log-location scale distribution. For convenience, we are going to assume that we are working with log-location scale distributions throughout the remainder of this section. The log-location scale distributions are a natural choice for survival/reliability data because they do not allow for negative failure times.

Although we are not going to cover location scale distributions here, going from a log-location scale distribution to a location scale distribution is trivial: Just carefully place some exp() and log() arguments. Let x denote failure time. If x has a log-location scale distribution, then log(x) has a location scale distribution. Table 2 provides a summary of the distributions currently available in JMP, using $\Phi_N(\cdot)$ to denote the cumulative distribution function of the standard normal distribution.

4.2. Constant censoring times

To make things easier, we assume a constant censor time for all observations. This assumption is appropriate in reliability studies because the reliability study is conducted for a fixed amount of time. Any units that have not failed by the end of the study are censored at the same time (c). The constant censoring time simplifies the likelihood function considerably

$$L(\mu,\sigma|X) = \Phi\left[\frac{\log(c)-\mu}{\sigma}\right]^{n_c} \prod_{x_i<c} \phi\left[\frac{\log(x_i)-\mu}{\sigma}\right], \tag{15}$$

where x_i are failure times, μ is the location parameter, σ is the scale parameter, n_c is the number of censored observations, and $\Phi(\cdot)$ and $\phi(\cdot)$ are the assumed standard cumulative distribution and probability density functions respectively.

The assumption of constant censoring times may not be appropriate outside of engineering studies. For example, in clinical trials a participant may be lost to follow-up at any point during the study.

4.3. What's up with the Weibull?

The Weibull distribution has two different parameterizations. Internally, we treat the Weibull distribution as a typical log-location scale distribution with probability density function

$$f(x|\mu,\sigma) = \frac{1}{x\sigma}\exp\left[\frac{\log(x)-\mu}{\sigma}\right]\exp\left\{-\exp\left[\frac{\log(x)-\mu}{\sigma}\right]\right\}$$

because that makes calculations much more convenient. Unfortunately, the 'α,β' parameterization of the Weibull distribution is more familiar to many engineers

$$f(x|\mu,\sigma) = \frac{\beta}{\alpha^\beta}x^{\beta-1}\exp\left[-\left(\frac{x}{\alpha}\right)^\beta\right]. \tag{16}$$

So we convert from the 'α, β' parameterization to the location scale parameterization using the following relationship

$$\mu = \log(\alpha) \text{ and } \sigma = \frac{1}{\beta}.$$

When we report the covariance matrix of the Weibull parameters, we use the location scale parameters since we use those parameters in further calculations.

5. Reliability Demonstration Plans

The goal of a reliability demonstration is to show that a new product meets or exceeds some reliability standard. For example, suppose that we are producing cars with a GPS system in the dash. We currently use GPS system A, which we have found to have a 90% probability of surviving one year of use. But GPS system B is less expensive, so we are considering putting it in our cars but are not willing to sacrifice reliability. We would use a reliability demonstration to find out if system B is at least as reliable as system A.

The first step in planning the demonstration is assuming a failure time distribution (Weibull, log-normal, ...) and a scale parameter σ. Often in these situations, engineers have a good idea about the distribution and scale either from prior experience or expert opinion. After specifying these quantities, the GPS reliability standard can be written

$$.9 = 1 - \Phi\left[\frac{\log(1) - \mu}{\sigma^\star}\right]$$

where $\Phi(\cdot)$ is the CDF of the assumed failure time distribution and $\sigma^\star$ is the assumed scale parameter. More generally, the reliability standard is stated

$$p^\star = 1 - \Phi\left[\frac{\log t^\star - \mu^\star}{\sigma^\star}\right] \tag{17}$$

where $p^\star$ is the standard probability of survival at time $t^\star$ and $\mu^\star$ is the location parameter associated with the reliability standard. Typically $\mu^\star$ is not explicitly stated; instead it is solved for using

$$\mu^\star = \log(t^\star) - \sigma^\star \Phi^{-1}\left(1 - p^\star\right). \tag{18}$$

By stating the reliability standard as in (17), we are able to solve for the location parameter of the assumed reliability standard using (18).

5.1. Reliability standard as a hypothesis test

In order to be able to formulate a demonstration plan, we pose the reliability demonstration as a hypothesis test. We do that by testing

$$\mathrm{H}_0 : p < p^\star \text{ vs } \mathrm{H}_a : p \geq p^\star \tag{19}$$

where p is the probability of survival at time $t^\star$ for the new product. If $p \geq p^\star$, then the new product is at least as reliable as the standard and we would like to reject the null hypothesis with high probability. When the test rejects, we say that the new product passed the reliability demonstration.

So how do we decide when to reject the null hypothesis? A natural choice is to test n units for t units of time. If one or more units fail before time t, we do not reject the null hypothesis. If all n units survive to time t, then we reject the null hypothesis and conclude that the product is at least as reliable as the standard.

More generally, we can say that we are willing to tolerate k failures during the demonstration. If more than k units fail before time t, then we do not reject the null hypothesis. If k or fewer units fail, then we reject the null hypothesis and conclude that the new product is at least as reliable as the standard. Choosing $k = 0$ will give us the quickest/cheapest experiment. Choosing larger k will result in a more expensive study in terms of time spent and units tested, but also more power to detect a small increase in reliability. We will see an example of this trade-off in Section 5.4.

The two quantities that define a demonstration are the number of units tested (n) and the length of time for testing them (t). If we define one of these quantities, we can solve for the other using information from the reliability standard and the assumed failure time distribution. In general n and t are inversely related, meaning that increasing the number of units tested will lower the amount of time needed to test them and vice versa.

5.2. Solving for the study length and the number of units tested

In the previous subsection we showed how to view a reliability demonstration as a hypothesis test. Now all that remains is to find the combination of sample size (n) and length of study (t) that will give the reliability study the desired false-positive rate (α). Since we intend to test n independent units in the demonstration, the number of failures has a binomial(n, p) distribution where p is equal to the probability of a unit failing before time t. We want to test (19) at the α level, so we know that

$$\begin{aligned} \text{Pr(reject H}_0 \mid \text{H}_0 \text{ true)} &= \text{Pr(} k \text{ or fewer failures} \mid \text{H}_0 \text{ true)} \\ \alpha &= \Pr\left(x \leq k \mid x \sim \text{Bin}\left\{n, \Phi\left[\frac{\log(t) - \mu^\star}{\sigma^\star}\right]\right\}\right) \end{aligned} \tag{20}$$

where $\mu^\star$ and $\sigma^\star$ come from the reliability standard. Plugging (18) into (20), we find that

$$\alpha = \Pr\left(x \leq k \mid x \sim \text{Bin}\left\{n, \Phi\left[\frac{\log(t) - \log(t^\star) + \sigma^\star\Phi^{-1}(1-p^\star)}{\sigma^\star}\right]\right\}\right), \tag{21}$$

which includes the two unknown values t and n. Given the length of the demonstration, we use (21) to solve for the number of units tested. Similarly, given the number of units tested, (21) will lead us to the required length of the demonstration.

Before going any further with (21), we must take advantage of some properties of the binomial and beta distributions. Recall that if x has a binomial(n,p) distribution,

$$\text{Pr}(x \leq k) = \text{I}_{1-p}(n-k, k+1)$$

where $\text{I}_x(a, b)$ is the regularized incomplete beta function

$$\text{I}_x(a,b) = \frac{\int_0^x t^{a-1}(1-t)^{b-1}\text{d}t}{\int_0^1 t^{a-1}(1-t)^{b-1}\text{d}t}$$

which can also be rewritten in terms of the gamma function. Writing the cumulative distribution function in this manner is useful because it provides a connection to the Beta distribution. Recall that

$$\text{Pr}(Y \leq y \mid Y \sim \text{Beta}(\alpha,\beta)) = \text{I}_y(\alpha,\beta) = \text{B}(y;\alpha,\beta).$$

This relationship will allow us to use C++ routines written for the Beta distribution, which will make calculations much more convenient.

Now using the properties of the binomial and beta distributions, we can solve for t given that there are n units to be tested. Continuing on from (21),

$$\begin{aligned}
\alpha &= \mathrm{I}_{1-\delta}(n-k, k+1) \\
&\qquad \text{where } \delta = \Phi\left[\frac{\log(t) - \log(t^\star) + \sigma^\star\Phi^{-1}(1-p^\star)}{\sigma^\star}\right] \\
\mathrm{B}^{-1}(\alpha; n-k, k+1) &= 1 - \Phi\left[\frac{\log(t) - \log(t^\star) + \sigma^\star\Phi^{-1}(1-p^\star)}{\sigma^\star}\right] \\
\log(t) - \log(t^\star) + \sigma^\star\Phi(1-p^\star) &= \sigma^\star\Phi^{-1}\left[1 - \mathrm{B}^{-1}(\alpha; n-k, k+1)\right] \\
t &= \frac{t^\star \text{exp}\left\{\sigma^\star\Phi^{-1}\left[1 - \mathrm{B}^{-1}(\alpha; n-k, k+1)\right]\right\}}{\text{exp}\left[\sigma^\star\Phi^{-1}(1-p^\star)\right]} \qquad (22)
\end{aligned}$$

where $\mathrm{B}^{-1}(\alpha; n-k, k+1)$ is the α quantile of the Beta$(n-k, k+1)$ distribution. Given a testing level, number of failures tolerated, number of units tested, and reliability standard, formula (22) provides a convenient solution for the necessary length of the demonstration.

Unfortunately, solving for n when given t is not quite as convenient. Now we are interested in finding the sample size n such that

$$\begin{aligned}
\alpha &= \mathrm{I}_{1-\delta}(n-k, k+1) \\
&\qquad \text{where } \delta = \Phi\left[\frac{\log(t) - \log(t^\star) + \sigma^\star\Phi^{-1}(1-p^\star)}{\sigma^\star}\right]
\end{aligned}$$

where t is known. This is equivalent to finding n that satisfies

$$\mathrm{B}^{-1}(\alpha; n-k, k+1) - 1 + \Phi\left[\frac{\log(t) - \log(t^\star) + \sigma^\star\Phi^{-1}(1-p^\star)}{\sigma^\star}\right] = 0. \qquad (23)$$

Given α, n, k, and the reliability standard, (23) is a function of only n. So we can find the n that satisfies (23) using standard root-finding methods. In JMP we use Brent's method to find the root of (23), giving us the required sample size given that the length of the demonstration is t.

5.3. Example Demonstration Plan

We will go back to the GPS example at the beginning of this section. We want to show that the new GPS system has at least a 90% chance of surviving after 12 months. We assume that the failure times have a lognormal distribution with scale parameter $\sigma = 3$ and that we only have $n = 50$ GPS units to test. The assumption on the scale parameter may come from prior experience or from an expert opinion. If a passing demonstration is defined by zero failures during the study, how long do we need to test the GPS units to assure that there is only a 5% chance of passing the demonstration when the new unit is not as reliable as the standard?

Given the requirements, we know that we are working with the lognormal distribution and that $\alpha = .05$, $t^\star = 12$, $p^\star = .9$, $\sigma^\star = 3$, $k = 0$, and $n = 50$. Plugging these values into (22), we get

$$\begin{aligned}
t &= \frac{12\text{exp}\left\{3\Phi^{-1}\left[1 - \mathrm{B}^{-1}(.05; 50, 1)\right]\right\}}{\text{exp}[3\Phi^{-1}(1-.9)]} \\
&= 5.044,
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. So the GPS units need to be tested for just over five months.

How would the demonstration plan change if we decided to define success as observing two or fewer failures during the study? Then $k = 2$ and the rest of the parameters stay the same, leading to

$$\begin{aligned} t &= \frac{12\text{exp}\left\{3\Phi^{-1}\left[1 - \text{B}^{-1}(.05; 50 - 2, 1 + 2)\right]\right\}}{\text{exp}[3\Phi^{-1}(1 - .9)]} \\ &= 16.673. \end{aligned}$$

So tolerating more failures means testing the 50 units for more than three times as long. The longer testing time means greater power to detect an increase in reliability, but may not be feasible due to cost or time constraints.

5.4. Determining power to reject the null hypothesis

After planning a reliability demonstration as described in the previous subsection, it is important to consider the power to detect an increase in reliability. That is, what is the probability of passing the demonstration given that the new product is at least as reliable as the standard? If the true reliability of the new product is greater than the standard, then we can express the true reliability at time $t^\star$ as

$$rp^\star = 1 - \Phi\left[\frac{\log(t^\star) - \mu_t}{\sigma^\star}\right]$$

where $0 \leq r \leq 1/p^\star$ is the true reliability ratio and μ_t is the true location parameter of the failure time distribution of the new product. When $r \geq 1$, then the true reliability is at least that of the standard. Then the true location parameter is

$$\mu_t = \log(t^\star) - \sigma^\star\Phi^{-1}(1 - rp^\star).$$

We can use the true location parameter to determine the probability of passing the demonstration given a change in reliability. Then

$$\begin{aligned} \text{power} &= \text{Pr(pass demonstration} \mid \text{true reliability at time } t^\star \text{ is } rp^\star) \\ &= \text{Pr}\left(x \leq k \mid x \sim \text{Bin}\left\{n, 1 - \Phi\left[\frac{\log(t) - \log(t^\star) + \sigma^\star\Phi^{-1}(1 - rp^\star)}{\sigma^\star}\right]\right\}\right) \qquad (24) \end{aligned}$$

is the formula for determining the probability of passing the demonstration given that the true reliability at time $t^\star$ is $rp^\star$.

For the GPS example in the previous section, we decided to test 50 units for 5.044 months (allowing for zero failures) to determine whether or not the reliability at one year is at least .9. What if the true reliability at one year is .99 for the new product? We would plug $r = 1.1$ into (24) and find that the probability of passing the demonstration is .7998. We also considered a demonstration plan that tested 50 units for 16.673 months and allowed for two failures during the demonstration. If the true reliability at one year is .99, then the probability of passing the demonstration is .9708. So extending the study and allowing for more failures provides a much greater probability of passing the demonstration. For the two demonstration plans considered for the GPS example, Figure 5 plots the probability of passing the demonstration as a function of the improvement in reliability.

5.5. Demo plan summary

The reliability demonstration is a straightforward way to compare the reliability of a product to a standard. The basic idea is not complicated. We want to test n units for t units of time and if k or fewer units fail, then the product passes the demonstration, and we conclude that it is at least as reliable as the standard. After making a few distributional assumptions and specifying the reliability standard, we can solve for the necessary length of the study analytically as stated in (22). We cannot solve for the necessary sample size analytically, but a root-finding algorithm (like Brent's method) will get the job done.

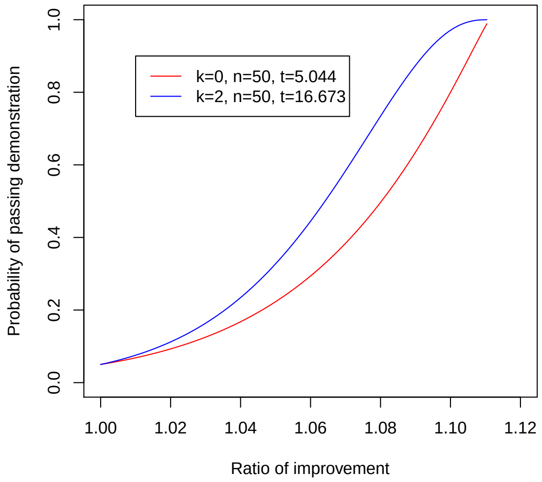

Fig. 5. Plot of power as a function of the reliability ratio of improvement

6. Reliability Test Plans

Suppose that we are interested in estimating the probability of a GPS unit failing before it has been in use for one year. Similarly, we may be interested in the time at which 50% of the units will have failed. In either situation, we would want to achieve some level of precision (based on confidence limits) about the estimate. The reliability test plan is designed to solve this problem. Given some distributional assumptions, it provides the sample size or length of study necessary to be able to estimate either a quantile or failure probability with the desired level of precision.

6.1. What is the objective?

In reliability tests, the objective is to either estimate the p^{th} quantile or the cumulative distribution function evaluated at time t. More explicitly, we are interested in

$$q(p) = \exp\left[\sigma\Phi^{-1}(p) + \mu\right]$$

or

$$p(t) = \Phi\left[\frac{\log(t) - \mu}{\sigma}\right]$$

where μ and σ are the location and scale parameters, respectively, and $\Phi(\cdot)$ is the standard cumulative distribution function of the assumed failure time distribution.

In order to define precision about $\hat{q}(p)$ and $\hat{p}(t)$ (which we will do in the following subsection), we need to have confidence intervals about the two quantities. Either way, we first have to find the expected information matrix for (μ, σ). As usual, we obtain the information matrix by taking the first two derivatives of (15). To get the expected information matrix, we use numerical integration tools that exist internally in JMP. From there we can invert the expected information matrix to get the expected covariance matrix Σ. Since we do not necessarily know the sample size, we calculate the expected information using a single observation and can later scale it by the defined sample size.

Once we have the expected information matrix, we can find the expected variance and Wald confidence interval (denoted $[c_l, c_u]$) for $\Phi^{-1}(p)\sigma + \mu$ using the delta method. Then we use $[e^{c_l}, e^{c_u}]$ as the expected confidence interval for $q(p)$. Note that this interval will not be symmetric because of the exponentiation. But the upper and lower endpoints of the interval are guaranteed to be non-negative, which is attractive because we know that log-location scale distributions are only defined for non-negative values.

For $p(t)$, first we use the delta method to find the expected variance of $z = \frac{\log(t)-\mu}{\sigma}$. From there we get the Wald confidence interval for z, denoted $[z_l, z_u]$. Then we use $[\Phi(z_l), \Phi(z_u)]$ as the expected confidence limits for $p(t)$. Once again, the confidence interval around $p(t)$ will not be symmetric because of the $\Phi(\cdot)$ transformation, but it is guaranteed to be contained in [0,1] (a favorable property for probabilities).

6.2. *Measures of precision*

In Section 6.1 we found ways to obtain expected confidence intervals. We will use $[q_l, q_u]$ to denote the confidence interval about $q(p)$ and $[p_l, p_u]$ to denote the confidence interval about $p(t)$. Now that the confidence intervals are defined, we define three different measures of precision based on them.

The *Interval Ratio* is simply the square root of the upper confidence limit divided by the square root of the lower confidence limit. For the quantile case, that results in $\sqrt{q_u/q_l}$. This precision definition allows us to say that the upper limit is only, say, 10% greater than the lower limit. Although this precision measure may not seem intuitive, it leads to convenient sample size formulas for quantiles that appear in Meeker and Escobar (1998). We know that the interval ratio will always be greater than one because the upper confidence limit has to be greater than or equal to the lower limit. Larger values of the interval ratio correspond to less precise measurements.

The *Two-sided Interval Absolute Width* is the width of the confidence interval. For example, the precision is $q_u - q_l$ in the case of the quantile. This precision measure allows us to say that we want the confidence interval around $q(p)$ to be only two, for example. Similarly, the *Lower One-sided Interval Absolute Width* is the width of the lower confidence limit. More specifically, if we are interested in a quantile, then the precision is defined as $q(p) - q_l$. The lower interval width is of interest because in some situations we may want to be able to say that "the .9 quantile is at least 100," for example. For these two measures, values closer to zero correspond to more precise measurements.

The last two precision measures can also be stated in a relative way. Instead of stating that you want the confidence interval around $q(.9)$ to be 1, we could say that we want the confidence interval around $q(.9)$ to be $.1 * q(.9)$. In other words, we want the width of the confidence interval to be some percentage of the point estimate. JMP will allow for the precision to be stated using these relative measures as well. To keep the internal calculations relatively simple, we simply convert these relative precisions back into absolute precisions and continue as usual.

6.3. *Solving for sample size or length of study*

In the two previous subsections, we looked at objectives of the reliability test plan and ways to define the desired precision. Given the failure time distributional assumptions, desired precision, and objectives, we are interested in determining the sample size or the length of the test. Given the length of the test, we can solve for the required sample size. Given the sample size, we can solve for required length of the test.

Given the length of the study is t, we can find the expected information matrix for (μ, σ) using numerical integration techniques that already exist in JMP. Then a naive way to find the sample size would be to start at say $n = 3$ and try out different sample sizes until we reach the desired precision level. Calculating the precision for each n is straightforward, using n in conjunction with the expected information matrix gives us the expected covariance matrix for (μ, σ). Then we use the expected covariance matrix to get an expected confidence interval, which leads to the precision calculation. Naturally, as n increases, we get more precise estimates. Rather than using this brute force to find n, we use Brent's method, which is much more efficient for finding roots.

Similarly, if we are given n, we could start at say $t = .1$ and increase t until we reach the desired precision. Once we get to the desired precision level, we stop and declare that $t^\star$ is the necessary study length. Rather than using this brute force approach, we use Brent's method to go about finding $t^\star$ much more efficiently. But there is a bump in the road when solving for t. For certain combinations of distributional assumptions, sample size, and desired precision, a sufficient study length may not exist. For example, if we are only testing $n = 5$ units but want an extremely precise measurement of the .5 quantile, even if we never stopped the study (no censoring, so $t = \infty$) we would not be able to achieve the desired precision. When this happens, the only option is to reconsider the distributional assumptions, increase the sample size, or choose a more realistic precision.

So for most of the situations that we will encounter with reliability test plans, we have to use numerical techniques to do the calculations. However, there are several situations where we can find an analytical solution for either n or t. The following subsections provide the details.

6.3.1. Sample size calculations for estimated quantiles

Luckily when we want to calculate the n that gives us precision δ for estimating q_p, we have analytical expressions for each definition of precision.

For the sample size that yields interval ratio δ, we solve for n using

$$\begin{aligned}
\delta^2 &= \frac{\text{exp}\left[\sigma\Phi^{-1}(p) + \mu + \sqrt{v/n}z_{1-\alpha/2}\right]}{\text{exp}\left[\sigma\Phi^{-1}(p) + \mu - \sqrt{v/n}z_{1-\alpha/2}\right]} \\
\delta^2 &= \text{exp}\left(2z_{1-\alpha/2}\sqrt{v/n}\right) \\
2\text{log}(\delta) &= 2z_{1-\alpha/2}\sqrt{v/n} \\
\sqrt{n} &= \frac{z_{1-\alpha/2}\sqrt{v}}{\text{log}(\delta)} \\
n &= \frac{z^2_{1-\alpha/2}v}{[\text{log}(\delta)]^2} \qquad (25)
\end{aligned}$$

where v is the expected variance of $\sigma\Phi^{-1}(p) + \mu$ when there is a single observation, α is the confidence level, and $z_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile of the standard normal distribution.

For the sample size that yields lower tail precision δ, we solve for n using

$$\begin{aligned}
\delta &= \text{exp}\left[\sigma\Phi^{-1}(p) + \mu\right] - \text{exp}\left[\sigma\Phi^{-1}(p) + \mu - \sqrt{v/n}z_{1-\alpha/2}\right] \\
\delta &= \gamma - \gamma\text{exp}\left(-z_{1-\alpha/2}\sqrt{v/n}\right) \\
\frac{\gamma}{\gamma - \delta} &= \text{exp}(z_{1-\alpha/2}\sqrt{v/n}) \\
n &= v\left[\frac{z_{1-\alpha/2}}{\text{log}(\gamma) - \text{log}(\gamma - \delta)}\right]^2
\end{aligned}$$

where $z_{1-\alpha/2}$, α, and v are defined as in (25) and $\gamma = \text{exp}(\sigma\Phi^{-1}(p) + \mu)$.

For the sample size that provides interval width equal to δ, the simplest strategy is to convert it back into an interval ratio problem. First recall that for interval ratio R,

$$R^2 = \text{exp}(2z_{1-\alpha/2}\sqrt{v/n})$$

and therefore

$$\log(R) = z_{1-\alpha/2}\sqrt{v/n}.$$

Then we can convert from the interval width precision measure to the interval ratio precision measure using

$$\begin{aligned}
\delta &= \exp\left[\log(\gamma) + \log(R)\right] - \exp\left[\log(\gamma) - \log(R)\right] \\
\delta &= \gamma R - \frac{\gamma}{R} \\
\frac{\delta}{\gamma} &= R - \frac{1}{R} \\
0 &= R^2 - R\frac{\delta}{\gamma} - 1
\end{aligned}$$

which is a prime candidate for being solved by the quadratic formula. Using the quadratic formula yields

$$R = \frac{\delta/\gamma + \sqrt{(\delta/\gamma)^2 + 4}}{2}. \tag{26}$$

We know that $R = (\delta/\gamma - \sqrt{(\delta/\gamma)^2 + 4})/2$ is not a useful solution because we know that the ratio cannot be negative. So if we want to find the sample size that yields expected interval width δ, we can use (26) to convert the problem into an interval ratio problem and use (25) to solve for sample size.

6.3.2. Sample size for estimated failure probabilities

When our goal is to estimate the failure probability $p(t)$, we are only able to find an analytical expression for n when we are using lower interval tail width as the precision measure. The expected Wald confidence interval for $p(t)$ is

$$\left[\Phi\left(\gamma - z_{1-\alpha/2}\sqrt{v/n}\right), \Phi\left(\gamma + z_{1-\alpha/2}\sqrt{v/n}\right)\right]$$

where $\gamma = (\log(t) - \mu)/\sigma$ and v is the expected variance of γ obtained using the delta method on the expected information matrix. Then to find the sample size needed for lower tail length δ, we solve

$$\begin{aligned}
\delta &= \Phi(\gamma) - \Phi\left(\gamma - z_{1-\alpha/2}\sqrt{v/n}\right) \\
\gamma - z_{1-\alpha/2}\sqrt{v/n} &= \Phi^{-1}\left[\Phi(\gamma) - \delta\right] \\
z_{1-\alpha/2}\sqrt{v/n} &= \gamma - \Phi^{-1}\left[\Phi(\gamma) - \delta\right] \\
n &= \frac{v z_{1-\alpha/2}^2}{\left[\delta - \Phi^{-1}(\gamma - \delta\right]^2}.
\end{aligned}$$

This formula provides the sample size needed to have expected lower tail width δ.

6.4. Example reliability test plan

Suppose we work for a battery company and are interested in estimating the median failure time of our AA batteries. We are going to assume that the failure time (in months) of these batteries has a lognormal distribution with location parameter $\mu = 2$ and scale parameter $\sigma = .75$. Figure 6 shows the assumed cumulative distribution function and median value at approximately $t = 7.389$. We can only wait five months for the test to run and we want the 95% confidence interval around the median to be approximately $\delta = 2$. So we need to determine the sample size needed to fit these requirements.

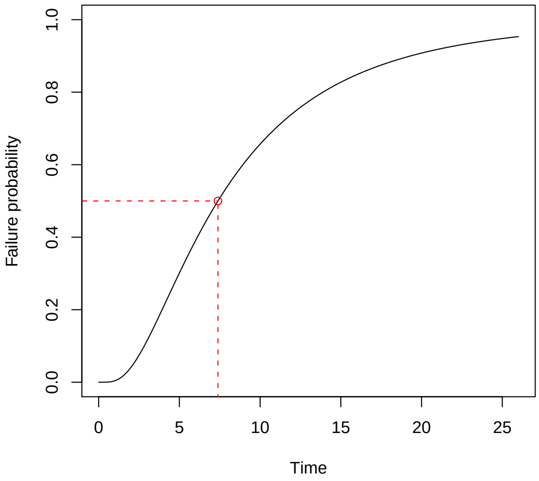

Fig. 6. Plot of the assumed failure time probability for the battery example. Median failure time is highlighted.

First we must convert the problem into an interval ratio problem using (26):

$$R = \frac{2/7.389 + \sqrt{(2/7.389)^2 + 4}}{2} \approx 1.144.$$

Then we can use $\delta = 1.144$ in (25):

$$n = \frac{1.96^2 v}{[\log(1.144)]^2} = 356.13$$

where $v = 1.6877$ is the approximate variance of the median obtained by using the delta method with the expected covariance matrix. Then we round up and say that testing $n = 357$ batteries for five months will yield the desired precision about the median failure time.

6.5. Some important diagnostics

The goal of the reliability test is to make inferences about a certain quantile or failure probability. We make these inferences based on the fitted failure time distribution. But in order to fit a distribution to our data, we must have enough information to be able to estimate all of the parameters. Since we use location scale distributions in JMP, that means that we have two parameters to estimate. And that means that we need to observe at least three failures during the reliability test to be able to safely estimate the model parameters. So when planning the reliability test, it is crucial to consider the number of failures that are expected during the test. If there are no failures during the test, the test will have been a waste of our resources.

In JMP we use two diagnostics to help us decide whether or not the test is going to be worthwhile. First we report the expected number of failures during the study

$$\begin{aligned} \text{E(number of failures)} &= n\text{Pr(unit fails before time } t_s) \\ &= n\Phi\left[\frac{\log(t_s) - \mu}{\sigma}\right], \end{aligned}$$

where t_s is the length of the study, $\Phi(\cdot)$ is the assumed CDF, n is the number of units being tested, and μ and σ are the assumed location and scale parameters, respectively. We also provide the user with the probability of observing fewer than three failures during the test:

$$\text{Pr(fewer than three failures)} = \text{Pr}\left\{x \leq 2 \mid x \sim \text{Bin}\left[n, \Phi\left(\frac{\log(t_s) - \mu}{\sigma}\right)\right]\right\}.$$

These two diagnostics are very useful in determining whether or not the test is worth pursuing.

For example, suppose again that we are making batteries and that we are interested in the median failure time. We assume (from past experience) that the failure time in months is lognormally distributed with $\mu = 1.25$ and $\sigma = 1.5$. Unfortunately because of cost constraints, we only want to test 20 batteries for one month. Then we compute our diagnostics

$$\begin{aligned}\text{E(number of failures)} &= 20\Phi\left[\frac{\log(1) - 1.25}{1.5}\right] \\ &\approx 4.05\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the standard normal CDF. And the probability of observing fewer than three failures is

$$\begin{aligned}\text{Pr(fewer than three failures)} &= \text{Pr}\left\{x \leq 2 \mid x \sim \text{Bin}\left[20, \Phi\left(\frac{\log(1) - 1.25}{1.5}\right)\right]\right\} \\ &\approx .20.\end{aligned}$$

So with only testing 20 batteries for one month, we expect to observe four failures, and there is a 20% chance that the test will not yield enough information to make any inferences. Given this information, we would probably decide not to risk wasting our time on this test and instead lobby for more resources.

6.6. Test plan summary

The reliability test involves testing a sample of units for a set period of time and recording failure times. Once the data are collected, we fit a distribution to the data and make inferences about failure probabilities or population quantiles. The test plan feature in JMP allows us to determine either the sample size or length of study necessary to achieve a desired level of precision about those inferences (based on the expected confidence interval). We must keep in mind, however, that for a given sample size, we can only reach a certain level of precision, even if we let all of the units run until failure. When planning a reliability test, it is also important to make sure that we expect to observe enough failures to be able to fit the assumed failure distribution.

7. Summary

The Power and Sample Size Calculator in JMP provides tools for planning experiments in a variety of settings. This document has covered some of the theory and resulting formulas that are used in the Power and Sample Size Calculator. The example JSL code and functions are useful for gaining better insight into these features.

References

[1] Agresti, A. and Caffo, B. (2000), “Simple and effective confidence intervals for proportions and differences of proportions result from adding two successes and two failures.” *The American Statistician*, 54, 280-288.

[2] Agresti, A. and Coull, B. (1998), “Approximate is better than ‘exact’ for interval estimation of binomial proportions.” *The American Statistician*, 52, 119-126.

[3] Meeker, W.Q. and Escobar, L. A. (1998), *Statistical Methods for Reliability Data*. Wiley: New York.

[4] National Institute of Standards and Technology (2009), “Engineering Statistics Handbook”. Gaithersburg: National Institute of Standards and Technology, http://www.itl.nist.gov/div898/handbook/ (accessed 10 June 2009).

A. Power and sample size formula cheat-sheet

This appendix provides a convenient summary of some of the formulas presented in the body of the paper. Throughout the appendix, we use α to denote the testing level 'Alpha'.

One Sample Mean
σ = Std Dev
p = Extra Parameters
δ = Difference to Detect
n = Sample Size

$$\text{Power} = 1 - F\left(f_{1-\alpha}, 1, n-p-1, \frac{n\delta^2}{\sigma^2}\right)$$

• $f_{1-\alpha}$ is the $1-\alpha$ quantile of the F(1,$n-p-1$) distribution.
• $F(x, \text{df}_1, \text{df}_2, nc)$ is the cumulative distribution function of the non-central F distribution with degrees of freedom df_1 and df_2 and non-centrality parameter nc evaluated at x.

Analytical solutions for δ and n do not exist, so we use numerical techniques in conjunction with the power formula to solve for them.

Two Sample Means
σ = Std Dev
p = Extra Parameters
δ = Difference to Detect
n = Sample Size per group

$$\text{Power} = 1 - F\left(f_{1-\alpha}, 1, 2n - p - 2, \frac{n\delta^2}{2\sigma^2}\right)$$

- $f_{1-\alpha}$ is the $1 - \alpha$ quantile of the F(1,$2n - p - 2$) distribution.
- $F(x, \text{df}_1, \text{df}_2, nc)$ is the cumulative distribution function of the non-central F distribution with degrees of freedom df_1 and df_2 and non-centrality parameter nc evaluated at x.

Analytical solutions for δ and n do no exist, so we use numerical techniques in conjunction with the power formula to solve for them.

k Sample Means
σ = Std Dev
p = Extra Parameters
$\mu_1, \mu_2, \ldots, \mu_k$: assumed mean for each of the k groups
n = Sample Size per group

$$\text{Power} = 1 - F\left(f_{1-\alpha}, k - 1, kn - p - k, \frac{n\sum_{j=1}^{k}(\mu_j - \bar{\mu})^2}{\sigma^2}\right)$$

- $\bar{\mu}$ = $\sum_{j=1}^{k} \mu_j / k$
- $f_{1-\alpha}$ is the $1 - \alpha$ quantile of the F(k-1,$kn - p - k$) distribution.
- $F(x, \text{df}_1, \text{df}_2, nc)$ is the cumulative distribution function of the non-central F distribution with degrees of freedom df_1 and df_2 and non-centrality parameter nc evaluated at x.

One Sample Standard Deviation
σ_0 = Hypothesized Standard Deviation
δ = Difference to Detect
n = Sample Size
$1-\beta$ = Power

Case 1: Alternative Standard Deviation Is Larger

$$\begin{aligned} 1-\beta &= 1-F\left(\frac{\sigma_0^2\chi_{1-\alpha}}{(\sigma_0+\delta)^2}\,,n-1\right) \\ \delta &= \sigma_0\sqrt{\frac{\chi_{1-\alpha}}{\chi_\beta}}-\sigma_0 \end{aligned}$$

Case 2: Alternative Standard Deviation Is Smaller

$$\begin{aligned} 1-\beta &= 1-F\left(\frac{\sigma_0^2\chi_{\alpha}}{(\sigma_0+\delta)^2}\,,n-1\right) \\ \delta &= \sigma_0\sqrt{\frac{\chi_{\alpha}}{\chi_{1-\beta}}}-\sigma_0 \end{aligned}$$

- χ_p is the p^{th} quantile of the χ^2_{n-1} distribution.
- $F(x,n-1)$ is the CDF of the χ^2_{n-1} distribution evaluated at x.

An analytical solution for n does not exist for either case, so numerical methods are used to solve for sample size.

One Sample Proportion
p = Proportion
p_0 = Null Proportion
n = Sample Size

Here we will show how to calculate power for the two-sided alternative. Calculations for the one-sided alternative are carried out in a similar manner.

$$\text{Power} = \sum_{y=0}^{n} \text{Pr}\left\{Y = y \,\middle|\, Y \sim \text{Binomial}(n, p)\right\} I\left\{T(y) \geq \chi_{1-\alpha}\right\}$$

where

$$T(y) = \frac{(\hat{p} - p_0)^2}{\frac{\hat{p}(1-\hat{p})}{n+4}} \text{ and } \hat{p} = \frac{y+2}{n+4}.$$

• $\chi_{1-\alpha}$ is the $1-\alpha$ quantile of the χ_1^2 distribution.
• $I\left\{T(y) \geq \chi_{1-\alpha}\right\}$ equals one when $T(y) \geq \chi_{1-\alpha}$ and zero otherwise.

Here we do not have closed-form expressions to solve for n or p_0, so numerical techniques are used to solve for them.

Two Sample Proportions
p_1 = Proportion 1
p_2 = Proportion 2
δ_0 = Null Difference in Proportion
n_1 = Sample Size 1
n_2 = Sample Size 2

Here we will show how to calculate power for the two-sided alternative. Calculations for the one-sided alternative are carried out in a similar manner.

$$\text{Power} = \sum_{y_1=0}^{n_1} \sum_{y_2=0}^{n_2} \Pr\left(Y_1 = y_1\right) \Pr\left(Y_2 = y_2\right) I\left\{T(y_1, y_2) \geq \chi_{1-\alpha}\right\}$$

where

$$Y_1 \sim \text{Binomial}(n_1, p_1)\ , Y_2 \sim \text{Binomial}(n_2, p_2),$$

$$T(y_1, y_2) = \frac{\left(\hat{p}_1 - \hat{p}_2 - \delta_0\right)^2}{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1+2} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2+2}}, \text{ and } \hat{p_j} = \frac{y_j + 1}{n_j + 2} \text{ for } j = 1, 2.$$

- $\chi_{1-\alpha}$ is the $1 - \alpha$ quantile of the χ^2_1 distribution.
- $I\left\{T(y_1, y_2) \geq \chi_{1-\alpha}\right\}$ equals one when $T(y_1, y_2) \geq \chi_{1-\alpha}$ and zero otherwise.

Here we do not have closed-form expressions to solve for n or p_0, so numerical techniques are used to solve for them.

Counts per Unit

λ_0 = Baseline Count per Unit
δ = Difference to Detect
n = Sample Size
$1 - \beta$ = Power

$$\begin{aligned} \text{Power} &= 1 - \Phi\left(\frac{Z_{1-\alpha} - \delta\sqrt{n/\lambda_0}}{\sqrt{(\lambda_0 + \delta)/\lambda_0}}\right) \\ n &= \frac{\lambda_0}{\delta^2}\left(Z_{1-\alpha} - Z_{\beta}\sqrt{\frac{\lambda_0 + \delta}{\lambda_0}}\right)^2 \end{aligned}$$

- $\Phi()$ is the standard normal CDF.
- Z_{1-p} is the $1 - p$ quantile of the standard normal distribution.

An analytical solution for δ does not exist, so numerical methods must be used to solve for δ given power and n.

Sigma Quality Level

σq = Sigma Quality Level
n = number of opportunities
d = number of defects

$$\begin{aligned} \sigma\text{q} &= Z_{1-d/n} + 1.5 \\ d &= n\left[1 - \Phi\left(\sigma\text{q} - 1.5\right)\right] \\ n &= d\left[1 - \Phi\left(\sigma\text{q} - 1.5\right)\right]^{-1} \end{aligned}$$

- $\Phi()$ is the standard normal CDF.
- $Z_{1-d/n}$ is the $1 - d/n$ quantile of the standard normal distribution.