Problem Set on Acceptance Tests

List of problems:

1. **Sampling until the first failure** (p.7). We will sample a very large population of items and test each item for integrity. Each item will either pass (P) or fail (F). We will stop sampling when the first F occurs.

   (a) What is the probability distribution of the sample size? What is the average and standard deviation of the sample size if the fraction of F's in the population is $p = 0.01$?

   (b) We have sampled $N = 215$ items without finding an F. Test the null hypothesis:

   $$H_0 : \quad p = 0.01 \tag{1}$$

   against the alternative hypothesis:

   $$H_1 : \quad p < 0.01 \tag{2}$$

   (c) What is the smallest sample size at which we reject H_0 in eq.(1) at a 0.025 level of confidence, if no F has yet been obtained?

   (d) In a different batch the first F occurred at $N = 70$ items. Test the null hypothesis:

   $$H_0 : \quad p = 0.01 \tag{3}$$

   against the alternative hypothesis:

   $$H_1 : \quad p > 0.01 \tag{4}$$

   Repeat this test if the first failure occurs at $N = 10$, and again at $N = 2$. What do you conclude from this?

2. **Fixed sample size.** (p.8) We will sample N items from a very large population, where each item either passes (P) or fails (F). The value of N is chosen before we begin. The probability of an F is denoted p, but its value is unknown.

(a) What is the probability distribution of the number of F's among the N tests?

(b) If we observe m F's among the N tests then our estimate of p would be m/N. Specifically, if we had observed 1 F among the N tests our estimate of p would be $p_1 = 1/N$. However, suppose that we observe no F's among the N tests. We now wish to choose between the following two hypotheses:

$$H_0 : \quad p = p_1 \tag{5}$$

$$H_1 : \quad p < p_1 \tag{6}$$

(c) Suppose that p truly equals zero: F's do not and cannot occur. How many tests would be needed to demonstrate this? What is the implication, more generally, when testing for very small p?

3. **Test for equivalent performance.** The performance of a system is evaluated as 'Low', 'Moderate' or 'High'. The performances of two versions of this system are summarized in table 1. Do these two systems have equivalent performance?

Version	Low	Moderate	High	Total
I	160	140	40	340
II	40	60	60	160
Totals	200	200	100	500

Table 1: Data for problem 3.

4. **Run-time faults or programming errors?** A large computer program has been run repeatedly with different input streams. Inputs occur from many different sources and at various different times during operation. Run-time faults can arise either from input errors or from programming errors. In the former case we expect a Poisson distribution in time of run-time faults. In the latter case we expect fairly consistent and recurrent run-time faults, though not completely so since some inputs may not activate some programming faults.

   13 runs completed without any run-time faults; 13 runs incurred a single fault; 4 runs incurred two faults each, and 2 runs had 3 faults. These data are summarized in table 2.

# of faults/run	0	1	2	3	
# of runs	13	13	4	2	Total = 32
# of faults	0	13	8	6	Total = 27

Table 2: Data for problem 4.

   With what confidence do you accept or reject the contention that the faults arose from input errors? (Hint: Estimate the coefficient of the Poisson distribution from the data, which removes an additional degree of freedom.)

5. **Milled surface.** The thickness of a milled surface is assessed as the deviation above a reference level. This deviation is measured in microns using a profilometer. In a particular sample the deviation was measured 8 times at a single point, with the following results: 12.30, 12.37, 12.37, 11.79, 12.17, 11.86, 12.31, 11.99. The desired deviation above the reference level is

11.90 microns. Assuming that the measurement errors are normally distributed, is this milled surface acceptably smooth?

6. **Thermal expansion.** The coefficient of thermal expansion of two welded metals must match, in order to prevent cracking during thermal cycling. The expansion coefficients are measured for random samples of the two metals, with the following results, in units of 10^{-6}/C (strain/degree C):

   Metal 1: 21, 24, 22, 23, 25, 22

   Metal 2: 24, 23, 22, 25, 25, 24, 25, 23

   Would you recommend welding these two metals for a thermal-cycling application? (Assume normal distribution of errors.)

7. **Vibration amplitude.** The amplitude of vibration in a milling machine increases as the tool bit wears out. An automatic monitoring system detects increased vibration in order to warn the operator of tool bit wear. Vibration levels in the milling machine are automatically categorized by the monitoring system as 'low', 'moderate' or 'high'. The vibration level is sampled periodically. During normal use of a sharp new bit, the fraction of 'low', 'moderate' and 'high' vibrations are $p_1 = 0.85$, $p_2 = 0.10$ and $p_3 = 0.05$, respectively.

   Of the past 100 samples, 72 samples are at the 'low' vibration level, 17 are 'moderate' and 11 samples are 'high'. Should the operator change the tool bit?

8. **Dispute about the mean.** One person claims that the random variable x is uniformly distributed on the interval $[0,\ 1]$. Another person claims that x has an average greater than 1/2. A large random sample of size $N = 50$ has been made, and the sample mean is $\overline{x} = 0.53$. Construct and implement a statistical hypothesis test to test the first person's claim as opposed to the second person's claim.

9. **Robustness of the significance level, with info-gap-uncertain pdf.** (p.13) Consider a test between two hypotheses, based on a sample of size n. Under hypothesis H_0 the sample mean is thought to be normally distributed with known mean and variance, μ_0 and σ^2/n. Under H_1 the distribution of the sample mean is shifted to the right by a known positive quantity, δ. We wish to distinguish between the hypotheses:

$$H_0: \quad \mu = \mu_0 \tag{7}$$

$$H_1: \quad \mu = \mu_0 + \delta \tag{8}$$

   While the actual distribution under H_1 is known to be the distribution under H_0 shifted by δ, the actual shape of the distribution in both cases is uncertain. Let $\widetilde{f}(\overline{x})$ denote the best estimate of the pdf of the sample mean under H_0, which is $\mathcal{N}(\mu_0, \sigma^2/n)$. An info-gap model for uncertainty in the actual distribution under H_0 is:

$$\mathcal{U}(h, \widetilde{f}) = \left\{ f(\overline{x}):\ f(\overline{x}) \geq 0,\ \int_{-\infty}^{\infty} f(\overline{x})\, \mathrm{d}\overline{x} = 1,\ |f(\overline{x}) - \widetilde{f}(\overline{x})| \leq h\widetilde{f}(\overline{x}),\ \forall \overline{x} \right\}, \quad h \geq 0 \tag{9}$$

   The critical value, C, is the rejection threshold: reject H_0 if and only if $\overline{x} > C$. The significance level, given pdf $f(\overline{x})$, is the probability of falsely rejecting H_0:

$$SL(f) = \mathsf{Prob}(\overline{x} > C | H_0) \tag{10}$$

$$= \int_C^{\infty} f(\overline{x})\, \mathrm{d}\overline{x} \tag{11}$$

The robustness of the test, for significance level α, is the greatest horizon of uncertainty in the pdf, up to which the significance level does not exceed α. Derive an explicit expression for the robustness, for small α.

10. **Robustness of the t test to distributional uncertainty.** (p.13) Use matlab program du03.m[1] to explore several properties of the robustnesses to distributional uncertainty for type I and type II errors when using a t test.

    (a) The robustnesses to distributional uncertainty for type I and type II errors are $\widehat{h}_0(t, \alpha^\star, \alpha)$ and $\widehat{h}_1(t, \alpha^\star, \beta)$. Explore how these robustnesses vary with variation of the number of degrees of freedom of the t test. Explain your results intuitively.

    (b) Level of significance, α, and power, $1 - \beta$, trade off against each other. For instance, for an ordinary t test with 17 DoFs, levels of significance of 0.01, 0.03 and 0.05 have associated powers of 0.15, 0.31 and 0.41, respectively. Explore this trade-off at different levels of robustness. Specifically, when:

$$\widehat{h}_0(t, \alpha^\star, \alpha) = \widehat{h}_1(t, \alpha^\star, \beta) = \text{constant} \tag{12}$$

    what are the values for α and $1 - \beta$ for $\alpha^\star = 0.01$, 0.03 and 0.05?

11. **Maximum likelihood estimate.** (p.14). Given a random sample, $x = (x_1, \ldots, x_n)$, what is the maximum likelihood estimate of the parameter λ for each of the following distributions:

    (a) Uniform distribution:

$$p(x) = \begin{cases} \frac{1}{\lambda}, & \text{if } 0 \le x \le \lambda \\ 0, & \text{else} \end{cases} \tag{13}$$

    (b) Triangular distribution:

$$p(x) = \begin{cases} -\frac{2x}{\lambda^2} + \frac{2}{\lambda}, & \text{if } 0 \le x \le \lambda \\ 0, & \text{else} \end{cases} \tag{14}$$

    (c) Exponential distribution:

$$p(x) = \lambda \mathrm{e}^{-\lambda x},\ x \ge 0 \tag{15}$$

12. **Accelerated lifetime testing: simple case,**[2] p.4 (p.16). The lifetime of a device is denoted ℓ, which depends on the "stress", s, which the system is subject to: $\ell(s)$. We know that the lifetime is zero for any stress exceeding the value s_0. Also, it is believed that the lifetime-stress relation is roughly linear for lower stress:

$$\ell_{\mathrm{m}}(s, c) = \begin{cases} (s - s_0)c & \text{if } s \le s_0 \\ 0 & \text{if } s \ge s_0 \end{cases} \tag{16}$$

    where $c < 0$.

    We have measured the lifetime, $\ell(s_1)$, at stress $s_1 < s_0$. We wish to estimate the lifetime at lower stress $s_2 < s_1$. Consider the following specific case. $s_0 = 1$, $s_1 = 0.8$, $\ell(s_1) = 10$ and $s_2 = 0.4$.

[1] \papers\T-Test\du03.m

[2] See Lecture Notes on Acceptance Testing, section 10 (acctes.tex).

(a) Use the known lifetimes to evaluate the coefficient $\widehat{c}$ in eq.(16) and to predict the lifetime at stress s_2.

(b) Our understanding indicates that the lifetime at low stress s_2 will be longer than the lifetime which is predicted based on the measurement at stress s_1. However, we do not know how much longer. Using the estimated coefficient $\widehat{c}$, calculate the robustness to lifetime uncertainty at critical errors 0, 0.2, 0.35 and 0.65. Discuss the meaning of these results.

(c) Now consider a more negative coefficient (steeper slope) $c = 1.01\widehat{c}$. What is the lifetime prediction with this value of c? Calculate the robustness for this c at the critical errors 0, 0.2, 0.35 and 0.65. Discuss the meaning of these results.

(d) Now consider an even more negative coefficient (steeper slope) $c = 1.02\widehat{c}$. What is the lifetime prediction with this value of c? Calculate the robustness for this c at the critical errors 0, 0.2, 0.35 and 0.65. Discuss the meaning of these results.

13. **Single hypothesis test,** p.5 (p.18). A particular property, x, (e.g. height, temperature, longevity, etc.) of a healthy population is a random variable with known mean and variance μ and σ^2. The central limit theorem asserts that, regardless of the distribution of x, the mean, $\overline{x}$, of a large random sample of size N is normally distributed with mean μ and variance σ^2/N. Given an observed value of the sample mean, $\overline{x}_{\rm obs}$, we want to test the hypothesis that the population is healthy:

$$H_0: \quad \overline{x} \sim \mathcal{N}(\mu, \sigma^2/N) \tag{17}$$

We will reject H_0 if $\overline{x}_{\rm obs}$ is an implausible value, conditioned on H_0. Specifically, we reject H_0 if, conditioned on H_0:

$$\text{Prob}\left(\left| \frac{\overline{x}-\mu}{\sigma/\sqrt{N}} \right| > \left| \frac{\overline{x}_{\rm obs}-\mu}{\sigma/\sqrt{N}} \right| ;\ H_0 \right) \le \alpha \tag{18}$$

where α is a 'level of significance'. If H_0 holds and if the sample is statistically random, then the distribution of $\frac{\overline{x}-\mu}{\sigma/\sqrt{N}}$ would be standard normal, $\mathcal{N}(0,1)$. Let $\Phi(z)$ denote the cumulative probability distribution (CPD) for $\mathcal{N}(0,1)$. The problem is that we are unsure that the sample is truly random: statistically independent measurements from the same population. Thus we are unsure that the true distribution of $\frac{\overline{x}-\mu}{\sigma/\sqrt{N}}$, call it $F(\cdot)$, is actually $\Phi(\cdot)$. We represent this uncertainty with the following info-gap model in which we introduce a simplifying assumption that the distributions are symmetric around the origin:

$$\mathcal{U}(h) = \left\{ F(z):\ F(-\infty)=0,\ F(\infty)=1,\ F(z)=1-F(-z),\ \frac{\mathrm{d}F}{\mathrm{d}z} \ge 0,\ |F(z)-\Phi(z)| \le h \right\}, \quad h \ge 0 \tag{19}$$

(a) We have an observed value of the sample mean, $\overline{x}_{\rm obs}$. Suppose that eq.(18) holds based on this observation, implying that we should reject H_0. How much can the distribution of $\frac{\overline{x}-\mu}{\sigma/\sqrt{N}}$ deviate from $\mathcal{N}(0,1)$ without changing this decision? That is, derive the robustness as a function of the rejection threshold, α.

(b) In contrast to part 13a, suppose that we have observed a sample mean, $\overline{x}_{\rm obs}$, for which eq.(18) does not hold, implying that we should accept H_0. Derive the robustness as a function of the rejection threshold, α.

Solutions to Homework on Acceptance Tests

List of problem solutions:

Solution to problem 1.

1a $x =$ number of samples to the first failure, and has a geometric distribution, where p is the probability of an F:

$$f(x) = (1-p)^{x-1}p, \quad x = 1,\, 2,\, \ldots \tag{20}$$

Check normalization:

$$\sum_{x=1}^{\infty} f(x) = p\sum_{x=1}^{\infty}(1-p)^{x-1} = p\sum_{x=0}^{\infty}(1-p)^{x} = p\frac{1}{1-(1-p)} = 1 \tag{21}$$

The mean and variance of x are:

$$\mathrm{E}(x) = \frac{1}{p}, \quad \mathsf{var}(x) = \frac{1-p}{p^2} \tag{22}$$

So:

$$\mathrm{E}(x) = \frac{1}{p} = \frac{1}{0.01} = 100 \tag{23}$$

$$\sqrt{\mathsf{var}(x)} = \sqrt{\frac{1-p}{p^2}} \approx \frac{1}{p} = 100 \tag{24}$$

1b Our observation is $x = N = 215$ samples without observing an "F". A large x is evidence against H_0. The level of significance is the probability of an equally or more extreme result (impugning H_0) conditioned on H_0. Hence the level of significance is:

$$\alpha = \mathsf{Pr}[x \geq N|H_0] = \sum_{x=N}^{\infty} f(x) = p\sum_{x=N}^{\infty}(1-p)^{x-1} = p(1-p)^{N-1}\underbrace{\sum_{x=0}^{\infty}(1-p)^{x}}_{\frac{1}{1-(1-p)}=\frac{1}{p}} = (1-p)^{N-1} \tag{25}$$

Under H_0: $p = 0.01$ so $\alpha = (0.99)^{N-1} = (0.99)^{214} = 0.116$. This is not small so we do not (yet) reject H_0.

1c We require $\alpha = (0.99)^{N-1} = 0.025$ hence:

$$N = 1 + \frac{\ln 0.025}{\ln 0.99} = 368.04 \tag{26}$$

So, at $N = 369$ we reject H_0 at $\alpha = 0.025$ level of significance.

1d The first F occurred at item $N = 70$. A small x impugns H_0. (This is the reverse of part 1b because H_1 in eq.(4) is the reverse of H_1 in eq.(2).) Hence the level of significance is:

$$\alpha = \mathsf{Pr}[x \le N|H_0] = \sum_{x=1}^{N} f(x) = p\sum_{x=1}^{N}(1-p)^{x-1} = p\frac{(1-p)^N - 1}{(1-p) - 1} = 1 - (1-p)^N \tag{27}$$

Under H_0: $p = 0.01$ so $\alpha = 1 - (0.99)^N = 1 - (0.99)^{70} = 0.505$. This is not small so we do not reject H_0.

Repeat this with the first F at $N = 10$, and we find that the level of significance is $\alpha = 1 - (0.99)^N = 1 - (0.99)^{10} = 0.0956$. This is not small so we still do not reject H_0.

Repeat this with the first F at $N = 2$, and we find that the level of significance is $\alpha = 1-(0.99)^N = 1 - (0.99)^2 = 0.0199$. This is small (but certainly not tiny) so we reject H_0.

The general conclusion is that it is "difficult" to reject H_0. The "problem" is that the standard deviation is nearly equal to the mean, as seen in eqs.(23) and (24).

Solution to problem 2.

2a The probability of exactly m F's among N tests is the binomial distribution:

$$f(m|N) = \left(\begin{array}{c} N \\ m \end{array} \right) p^m (1-p)^{N-m} \tag{28}$$

where the binomial coefficient is:

$$\left(\begin{array}{c} N \\ m \end{array} \right) = \frac{N!}{m!(N-m)!} \tag{29}$$

The mean and variance of the binomial distribution are:

$$\mathrm{E}(m) = Np, \quad \mathsf{var}(m) = Np(1-p) \tag{30}$$

Thus, for small p, we see:

$$\frac{\sqrt{\mathsf{var}(m)}}{\mathrm{E}(m)} = \frac{\sqrt{Np(1-p)}}{Np} = \sqrt{\frac{(1-p)}{Np}} \approx \frac{1}{\sqrt{Np}} \tag{31}$$

For instance, this ratio equals 1 for $N = 100$ and $p = 0.01$. In other words, the binomial distribution tends to be wide.

2b The level of significance, α, is the probability of an equally or more extreme result than observed (less than or equal to 0). A small value of α impugns H_0. Thus, with $p_1 = 1/N$:

$$\alpha = \mathsf{Pr}(x = 0|H_0) = f(0|N) = \left(\begin{array}{c} N \\ 0 \end{array} \right) p_1^0 (1-p_1)^N = \left(1 - \frac{1}{N}\right)^N \tag{32}$$

See results in table 3. The confidence converges very slowly (and is not even monotonic at very large N) as the sample size rises.

2c If you are looking for something that isn't there, you will never find it. And even if p is positive but very very tiny, you are very very unlikely to ever find an F. The implication is that estimating p based on fitness tests is infeasible when p is tiny. The best one can hope for is to establish confidence that p is no greater than a specified value. We explored this in problem 1d. We saw that this also is often inconclusive. If p is very small then fitness testing should probably be augmented with (or replaced by) other methods of system analysis. For example, attempting to identify possible failure mechanisms, and verifying their rarity.

N	α
2	0.25
5	0.3277
10	0.3487
20	0.3585
30	0.3617
50	0.3642
100	0.3666
10^3	0.3677
10^6	0.3679
10^{15}	0.3682
∞	1

Table 3: Levels of significance, eq.(32).

Solution to problem 3. The systems are equivalent if the rows and columns of the data, table 1, are statistically independent. We will use the χ^2 test to test this hypothesis.

c = number of columns = 3.

r = number of rows = 2.

$p_{ij} =$ probability of an outcome in row i and column j. We can estimate this probability as:

$$p_{ij} = \frac{n_{ij}}{N} \tag{33}$$

where $n_{ij} =$ number of outcomes in row i and column j.

$p_{\bullet i} =$ probability of an outcome in column i.

$p_{j\bullet} =$ probability of an outcome in row j.

The hypothesis of statistical independence of rows and columns is:

$$H_0: \quad p_{ij} = p_{i\bullet} p_{\bullet j} \tag{34}$$

The alternative hypothesis is:

$$H_1: \quad H_0 \text{ is false} \tag{35}$$

$$p_{\bullet 1} = \frac{200}{500} = 0.4, \quad p_{\bullet 2} = \frac{200}{500} = 0.4, \quad p_{\bullet 3} = \frac{100}{500} = 0.2, \quad p_{1\bullet} = \frac{340}{500} = 0.68, \quad p_{2\bullet} = \frac{160}{500} = 0.32 \tag{36}$$

$$\chi^2 = \sum_{i=1}^{c} \sum_{j=1}^{r} \frac{\left(n_{ij} - N p_{\bullet i} p_{j\bullet}\right)^2}{N p_{\bullet i} p_{j\bullet}} = 49.63 \tag{37}$$

DOF = number of categories $-$ number of constraints:

$$\text{DOF } = (r-1)(c-1) = 2 \tag{38}$$

$$\chi^2_{2,0.005} = 10.6 \implies \Pr\left(\chi^2_2 \geq 10.6\right) = 0.005 \implies \Pr\left(\chi^2_2 \geq 49.63\right) \ll 0.005 \tag{39}$$

So we strongly reject H_0, with level of confidence greater than $1 - 0.005$.

Solution to problem 4. We must test the following hypotheses:

H_0: Faults are due to input errors.

H_1: Faults are due to programming errors.

These hypotheses are **roughly** equivalent to:

H_0: The number of faults per run is a Poisson random variable.

H_1: The number of faults per run is **not** a Poisson random variable.

If H_0 is true, then n, the number of faults per run, will be distributed as:

$$P_n = \frac{\mathrm{e}^{-\lambda}\lambda^n}{n!} \tag{40}$$

where $\lambda =$ average number of faults per run.

We do not know λ but we can estimate it from the data:

$$\lambda = \frac{27}{32} \approx 0.84 \tag{41}$$

We can test H_0 against H_1 with a χ^2 test.

We have four categories:

Faults per run	Probability conditioned on H_0
0	$P_0 = \mathrm{e}^{-\lambda} = 0.432$
1	$P_1 = \lambda\mathrm{e}^{-\lambda} = 0.363$
2	$P_2 = \frac{\lambda^2\mathrm{e}^{-\lambda}}{2!} = 0.152$
3 or more	$\sum_{n=3}^{\infty} P_n = 0.053$

Table 4: Categories for χ^2 test for problem 4.

DOF = number of categories - number of constraints:

$$4 - 2 = 2 \tag{42}$$

1 constraint: normalization. 1 constraint: estimate of λ.

Define:

$n_k =$ number of runs with k faults.

$N =$ total number of runs.

$P_k =$ probability of run with k faults.

So we calculate the χ^2 statistic as:

$$\chi^2 = \sum_{k=1}^{4} \frac{(n_k - NP_k)^2}{NP_k} = 0.422 \tag{43}$$

The level of significance in our test is α = probability of a more extreme result, conditioned upon H_0:

$$\begin{aligned} \alpha &= \Pr\left(\chi_2^2 \geq \chi_{\mathrm{obs}}^2 | H_0\right) && (44)\\ &= \Pr\left(\chi_2^2 \geq 0.422 | H_0\right) && (45)\\ &> 0.5 && (46) \end{aligned}$$

This is **not small,** so we cannot reject H_0.

Solution to problem 5. Use the t statistic to test between the following two hypotheses:

$$H_0: \quad x = 11.90 \tag{47}$$

$$H_1: \quad x \neq 11.90 \tag{48}$$

The data result in:

$$N = 8, \quad \overline{x} = 12.145, \quad s^2 = 0.054943, \quad s = 0.23440 \tag{49}$$

Thus the t statistic, conditioned on H_0, is:

$$t = \frac{\overline{x} - 11.90}{s/\sqrt{8}} \sim t_{(7)} \tag{50}$$

The observed value of the statistic is:

$$t_o = \frac{12.145 - 11.90}{0.2344/\sqrt{8}} = 2.956 \tag{51}$$

If t_o is "greatly different" from zero, then reject H_0. We use the level of confidence to evaluate the degree of deviation of t_o from zero:

$$\alpha = \mathsf{Pr}\left(|t| \geq |t_o| \,\middle|\, H_0\right) \tag{52}$$

$$= 2 \times \left[\mathsf{Pr}\left(t_{(7)} \geq 2.956\right)\right] \tag{53}$$

$$= 2 \times (1 - 0.99) = 0.02 \tag{54}$$

This is small so we reject H_0.

Solution to problem 6. Use the t test to compare two means. The null- and alternative hypotheses are:

$$H_0: \quad \mu_1 = \mu_2 \tag{55}$$

$$H_1: \quad \mu_1 \neq \mu_2 \tag{56}$$

The basic results are:

Metal 1: $N_1 = 6 \quad \overline{x}_1 = 22.833 \quad s_1^2 = 2.1666 \quad s_1 = 1.47194$
Metal 2: $N_2 = 8 \quad \overline{x}_2 = 23.875 \quad s_2^2 = 1.2679 \quad s_2 = 1.1260$

Define:

$$\Delta = \overline{x}_1 - \overline{x}_2 \tag{57}$$

for which the observed value is:

$$\Delta_o = -1.042 \tag{58}$$

Under hypothesis H_o:

$$\mathrm{E}(\Delta) = 0, \;\; \mathsf{var}(\Delta) = \mathsf{var}(\overline{x}_1) + \mathsf{var}(\overline{x}_2) \approx \frac{s_1^2}{N_1} + \frac{s_2^2}{N_2} = \sigma_\Delta^2 = 0.5196 \tag{59}$$

Thus $\sigma_\Delta = 0.7208$.

Since the errors have a normal distribution we conclude, under H_0, that:

$$\Delta \sim t_{(N_1+N_2-2)} \tag{60}$$

A large value of $|\Delta|$ is evidence against H_0. The level of significance is:

$$\alpha = \mathsf{Pr}\left(|\Delta| \geq |\Delta_o| \;\middle|\; H_0 \right) \tag{61}$$

$$= \mathsf{Pr}\left(\frac{|\Delta|}{\sigma_\Delta} \geq \frac{|\Delta_o|}{\sigma_\Delta} \;\middle|\; H_0 \right) \tag{62}$$

$$\approx 2 \times 0.08 = 0.16 \tag{63}$$

α is not small so we do not reject H_0.

Solution to problem 7. Use the χ^2 test with $K = 3$ categories. In normal operation the probabilities of outcomes in each of the three categories are:

$$p_1^o = 0.85, \quad p_2^o = 0.10, \quad p_3^o = 0.05 \tag{64}$$

We wish to test between the following hypotheses:

$$H_0: \qquad p_i = p_i^o, \; i = 1, \ldots, 3 \tag{65}$$

$$H_1: \qquad H_0 \text{ is false} \tag{66}$$

The χ^2 statistic is:

$$\chi^2 = \sum_{i=1}^{K} \frac{(N_i - N p_i^o)^2}{N p_i^o} \tag{67}$$

where N_i is the number of outcomes in category i and N is the total number of outcomes. This statistic is distributed as $\chi^2_{(3-1)}$ if H_0 holds.

The observed statistic is:

$$\chi_o^2 = 14.08 \tag{68}$$

The level of significance is:

$$\alpha = \mathsf{Pr}(\chi^2_{(2)} \geq \chi^2_o \Big| \, H_0) = \mathsf{Pr}(\chi^2_{(2)} \geq 14.08) \leq 1 - 0.999 = 0.001 \tag{69}$$

This is very small so we reject H_0 and change the tool bit.

Solution to problem 8. The two hypotheses are:

$$H_0 \qquad x \sim U[0,\ 1] \tag{70}$$

$$H_1 \qquad \mathrm{E}(x) > 1/2 \tag{71}$$

Under H_0 the mean and variance of x are 1/2 and 1/12. Since the sample is large, the central limit theorem implies that, under H_0, $\overline{x} \sim \mathcal{N}(\frac{1}{2}, \frac{1/12}{N})$. Thus the level of significance is:

$$\begin{aligned}
\alpha &= \text{Prob}(\overline{x} \geq \overline{x}_o | H_0) && (72)\\
&= \text{Prob}\left(\frac{\overline{x} - \mu_{\overline{x}}}{\sigma_{\overline{x}}} \geq \frac{\overline{x}_o - \mu_{\overline{x}}}{\sigma_{\overline{x}}} \,\middle|\, H_0 \right) && (73)\\
&= 1 - \Phi\left(\frac{\overline{x}_o - \mu_{\overline{x}}}{\sigma_{\overline{x}}} \right) && (74)\\
&= 1 - \Phi\left(\frac{0.53 - 0.5}{1/\sqrt{12 \times 50}} \right) = 1 - \Phi(0.735) = 1 - 0.7673 = 0.2327 && (75)
\end{aligned}$$

This is not large so we cannot reject H_0.

Solution to problem 9. (p.3) The robustness is the greatest horizon of uncertainty up to which all pdf's result in significance level no less than α:

$$\widehat{h}(\alpha) = \max \left\{ h : \left(\max_{f \in \mathcal{U}(h, \widetilde{f})} SL(f) \right) \leq \alpha \right\} \tag{76}$$

Let $M(h)$ denote the inner maximum in eq.(76). For small significance level, $\alpha \ll 1$, this maximum occurs when the upper tail is as fat as possible at horizon of uncertainty h:

$$f(\overline{x}) = (1 + h)\widetilde{f}(\overline{x}) \tag{77}$$

Thus:

$$M(h) = (1 + h) \int_C^\infty \widetilde{f}(\overline{x}) \mathrm{d}\overline{x} \tag{78}$$

The robustness is the greatest value of h for which:

$$M(h) = \alpha \tag{79}$$

Let $\widetilde{\alpha}$ denote the best estimate of the significance level:

$$\widetilde{\alpha} = \int_C^\infty \widetilde{f}(\overline{x}) \mathrm{d}\overline{x} \tag{80}$$

Combining eqs.(78)–(80) we find the robustness to be:

$$\widehat{h}(\alpha) = \frac{\alpha}{\widetilde{\alpha}} - 1 \tag{81}$$

or zero if this expression is negative. Note:

- The robustness depends on the sample size via $\widetilde{\alpha}$, which depends on the sample size via the estimated distribution $\widetilde{f}$.
- The robustness is zero for the estimated significance, $\widetilde{\alpha}$.
- The robustness is positive for the significance greater (lower confidence) than $\widetilde{\alpha}$.

Solution to problem 10. (p.4)

(a) The robustnesses to distributional uncertainty for type I and type II errors are $\widehat{h}_0(t, \alpha^\star, \alpha)$ and $\widehat{h}_1(t, \alpha^\star, \beta)$. The t test is constructed with the nominal pdfs for level of significance $\alpha^\star$. $\widehat{h}_0(t, \alpha^\star, \alpha)$

DoF	$\alpha^\star = 0.01$		$\alpha^\star = 0.03$		$\alpha^\star = 0.05$	
	α for	$1-\beta$ for	α for	$1-\beta$ for	α for	$1-\beta$ for
	$\widehat{h}_0 = 0.04$	$\widehat{h}_1 = 0.2$	$\widehat{h}_0 = 0.04$	$\widehat{h}_1 = 0.2$	$\widehat{h}_0 = 0.04$	$\widehat{h}_1 = 0.2$
5	0.1210	0	0.1273	0.0674	0.1411	0.1510
17	0.1004	0.0480	0.1117	0.1608	0.1274	0.2392
50	0.0956	0.0752	0.1080	0.1884	0.1240	0.2627

Table 5: Results for problem 10(a).

is the greatest horizon of uncertainty up to which the probability of falsely rejecting H_0 is no greater than α. $\widehat{h}_1(t, \alpha^\star, \beta)$ is the greatest horizon of uncertainty up to which the probability of falsely accepting H_0 is no greater than β. The "power" is defined as $1-\beta$, the probability of falsely rejecting H_1.

Table 5 shows values α for which $\widehat{h}_0(t, \alpha^\star, \alpha) = 0.04$, and values of $1-\beta$ for which $\widehat{h}_1(t, \alpha^\star, \beta) = 0.2$, for various degrees of freedom and various values of $\alpha^\star$. From this table we observe the following two facts:

1. At fixed $\alpha^\star$ (the nominal level of significance), $\widehat{h}_0(t, \alpha^\star, \alpha)$ reaches robustness of 0.04 at lower α (more significant rejection) as the DoF increases. This means that the robustness increases as the DoF increases.

2. At fixed $\alpha^\star$ (the nominal level of significance), $\widehat{h}_1(t, \alpha^\star, \beta)$ reaches robustness of 0.2 at larger $1-\beta$ (greater power) as the DoF increases. This means that the robustness increases as the DoF increases.

(b) We seek the values of α and $1-\beta$ at which:

$$\widehat{h}_0(t, \alpha^\star, \alpha) = \widehat{h}_1(t, \alpha^\star, \beta) = 0,\ 0.04, \text{ and } 0.08 \tag{82}$$

$\widehat{h}_0 = \widehat{h}_1$	$\alpha^\star$= 0.01	0.03	0.05	$\alpha^\star$= 0.01	0.03	0.05
0	$\alpha =$ 0.0100	0.0300	0.0500	$1-\beta =$ 0.1505	0.3066	0.4068
0.04	0.1004	0.1117	0.1274	0.1172	0.2646	0.3605
0.08	0.1789	0.1828	0.1942	0.0946	0.2333	0.3248

Table 6: Results for problem 10(b).

Results are shown in table 6 for 17 DoFs. We observe:

1. Robust level of significance (α) trades-off against robust power $(1-\beta)$ at all levels of robustness.

2. The trade-off is less severe at high robustness than at low robustness. For instance, at robustness of 0.04, the level of significance deteriorates less than at robustness of zero, for approximately the same relative improvement in power.

Solution to problem 11. (p.4). Define $s(x) = 1$ if $x \geq 0$ and zero otherwise.

(a) The likelihood function is the product of pdf's:

$$L(x|\lambda) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \frac{1}{\lambda} s(\lambda - x_i) s(x_i) \tag{83}$$

$$= \frac{1}{\lambda^n} \text{ if } x_i \in [0,\ \lambda]\ \forall\ i \tag{84}$$

and zero otherwise.

Thus the MLE is:

$$\widehat{\lambda} = \max_i x_i \tag{85}$$

The MLE is not defined if there is a negative measurement, which simply refutes this pdf.

(b) The likelihood function is the product of pdf's:

$$L(x|\lambda) = \prod_{i=1}^{n} p(x_i) = \prod_{i=1}^{n} \left(-\frac{2x_i}{\lambda^2} + \frac{2}{\lambda} \right) s(\lambda - x_i) s(x_i) \tag{86}$$

$$\frac{\partial L}{\partial \lambda} = \frac{2}{\lambda^2} \prod_{i=1}^{n} \left(\frac{2x_i}{\lambda} - 1 \right) s(\lambda - x_i) s(x_i) \tag{87}$$

Thus the MLE is for λ to equal $2x_i$ for some measurement i. Also, we require $\lambda \geq \max_i x_i$. This allows several possible solutions. Consider the 2nd derivative:

$$\frac{\partial^2 L}{\partial \lambda^2} = \frac{4}{\lambda^3} \prod_{i=1}^{n} \left(-\frac{3x_i}{\lambda} + 1 \right) s(\lambda - x_i) s(x_i) \tag{88}$$

For a maximum we require that an odd number of terms be negative.

The MLE is not defined if there is a negative measurement, which simply refutes this pdf.

For instance:

Suppose $n = 1$ where $0 \leq x_1$. Then $\lambda = 2x_1$ implies $L' = 0$ and $L'' < 0$. So this is the MLE.

Suppose $n = 2$ where $0 \leq x_1 \leq x_2$, where $-\frac{3x_1}{2x_2} + 1 > 0$. Then $\lambda = 2x_2$ implies $L' = 0$ and $L'' < 0$. So this is the MLE.

Suppose $n = 2$ where $0 \leq x_1 \leq x_2$, where $-\frac{3x_1}{2x_2} + 1 < 0$. Then $\lambda = 2x_2$ implies $L' = 0$ and $L'' > 0$. So this is *not* the MLE.

(c) The likelihood function is the product of pdf's:

$$L(x|\lambda) = \prod_{i=1}^{n} p(x_i) = \lambda^n \prod_{i=1}^{n} \mathrm{e}^{-\lambda x_i} s(x_i) = \lambda^n \mathrm{e}^{-\lambda X} \prod_{i=1}^{n} s(x_i) \tag{89}$$

where $X = \sum_{i=1}^{n} x_i$.

$$\frac{\partial L}{\partial \lambda} = \lambda^{n-1} \mathrm{e}^{-\lambda X} (n - \lambda X) \prod_{i=1}^{n} s(x_i) \tag{90}$$

Thus the MLE is:

$$\widehat{\lambda} = \frac{n}{X} \tag{91}$$

The MLE is not defined if there is a negative measurement, which simply refutes this pdf.

Solution to problem 12. (p.4).

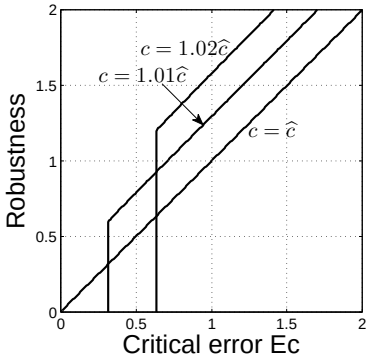

Figure 1: Robustness curves for problem 12.

(a) Given the data, we estimate c:

$$\widehat{c} = \frac{\ell(s_1)}{s_1 - s_0} = \frac{10}{0.8 - 1} = -50 \tag{92}$$

Thus the lifetime prediction is:

$$\ell(s_2) = (s_2 - s_0)\widehat{c} = (0.4 - 1)(-50) = 30 \tag{93}$$

(b) We know that the robustness curve for the estimated slope is a straight line at 45^o:

$$\widehat{h}(\widehat{c}, E_{\rm c}) = E_{\rm c} \tag{94}$$

Thus the robustness based on $\widehat{c}$ is shown in the 2nd column of table 7

$E_{\rm c}$	$\widehat{h}(\widehat{c},E_{\rm c})$	$\widehat{h}(1.01\widehat{c}, E_{\rm c})$	$\widehat{h}(1.02\widehat{c}, E_{\rm c})$
0	0	0	0
0.2	0.2	0	0
0.35	0.35	0.64	0
0.65	0.65	0.94	1.2

Table 7: Data for problem 3.

(c) The lifetime prediction with $c = 1.01\widehat{c} = -50.5$ is:

$$\ell(s_2) = (s_2 - s_0)c = (0.4 - 1)(-50.5) = 30.3 \tag{95}$$

A 1% increase in predicted lifetime.

To calculate the robustness we use:

$$\widehat{h}(c, E_{\rm c}) = \begin{cases} 0 & \text{if } E_{\rm c} \leq \sqrt{\mu_1} \\ \sqrt{E_{\rm c}^2 - [\ell(s_1) - \ell_{\rm m}(s_1, c)]^2} - \ell_{\rm m}(s_2, \widehat{c}) + \ell_{\rm m}(s_2, c) & \text{else} \end{cases} \tag{96}$$

where:

$$\mu_1 \quad = \quad [\ell(s_1) - \ell_{\rm m}(s_1, c)]^2 + [\ell_{\rm m}(s_2, \widehat{c}) - \ell_{\rm m}(s_2, c)]^2 \tag{97}$$

We find $\sqrt{\mu_1} = 0.32$, which is the value at which the robustness curve lifts off the E_c axis. Thus $\widehat{h}(c, 0) = \widehat{h}(c, 0.2) = 0$ while $\widehat{h}(c, 0.35)$ and $\widehat{h}(c, 0.65)$ will be positive:

$$\widehat{h}(c, 0.35) = \sqrt{0.35^2 - [10 - 10.1]^2} - 30.0 + 30.3 = 0.635 \quad (98)$$

$$\widehat{h}(c, 0.65) = \sqrt{0.65^2 - [10 - 10.1]^2} - 30.0 + 30.3 = 0.942 \quad (99)$$

See column 3 of table 7.

(d) The lifetime prediction with $c = 1.02\widehat{c} = -51$ is:

$$\ell(s_2) = (s_2 - s_0)c = (0.4 - 1)(-51) = 30.6 \quad (100)$$

A 2% increase in predicted lifetime.

We find $\sqrt{\mu_1} = 0.63$, which is the value at which the robustness curve lifts off the E_c axis. Thus $\widehat{h}(c, 0) = \widehat{h}(c, 0.2) = \widehat{h}(c, 0.35) = 0$ while $\widehat{h}(c, 0.65)$ will be positive:

$$\widehat{h}(c, 0.65) = \sqrt{0.65^2 - [10 - 10.2]^2} - 30.0 + 30.6 = 1.22 \quad (101)$$

See column 4 of table 7.

Solution to problem 13. (p.5).

(13a) Define the random variable $z = \frac{\overline{x}-\mu}{\sigma/\sqrt{N}}$ and denote its CPD as $F(z)$. Let $P_{\rm rej}(F)$ denote the probability of rejection in eq.(18) on p.5. Define $z_{\rm obs} = \frac{\overline{x}_{\rm obs}-\mu}{\sigma/\sqrt{N}}$. Thus:

$$P_{\rm rej}(F) = \text{Prob}(|z| \geq |z_{\rm obs}|) = F(-|z_{\rm obs}|) + [1 - F(|z_{\rm obs}|)] \quad (102)$$

$$= 2[1 - F(|z_{\rm obs}|)] \quad (103)$$

where eq.(103) exploits the symmetry around the origin of the CPD's in the info-gap model.

The definition of the robustness is:

$$\widehat{h}_1(\alpha) = \max\left\{h:\ \left(\max_{F\in\mathcal{U}(h)} 2[1 - F(|z_{\rm obs}|)]\right) \leq \alpha\right\} \quad (104)$$

Let $m(h)$ denote the inner maximum, which occurs when $F(|z_{\rm obs}|)$ is as small as possible at horizon of uncertainty h:

$$F(|z_{\rm obs}|) = \max\left[\frac{1}{2},\ \Phi(|z_{\rm obs}|) - h\right] \quad (105)$$

where the '$\frac{1}{2}$' comes from the symmetry of the CPD's in the info-gap model. Thus, for $h \leq \Phi(|z_{\rm obs}|) - \frac{1}{2}$:

$$m(h) = 2\left[1 - \Phi(|z_{\rm obs}|) + h\right] \leq \alpha \quad \Longrightarrow \quad \boxed{\widehat{h}_1(\alpha) = \frac{\alpha - 2[1 - \Phi(|z_{\rm obs}|)]}{2}} \quad (106)$$

or zero if this is negative.[3] Note that, for all $0 \leq \alpha \leq 1$, $\widehat{h}$ does not exceed $\Phi(|z_{\rm obs}|) - \frac{1}{2}$ so we need not evaluate $m(h)$ for larger values of h.

The robustness curve in eq.(106) is plotted in fig. 2 for the following parameter values: $\mu = 1$, $\sigma = 1.3$ and $N = 30$. The observed sample mean is $\overline{x}_{\rm obs} = 1.5$ for which $z_{\rm obs} = 2.1066$ and $\Phi(|z_{\rm obs}|) = 0.9824$.

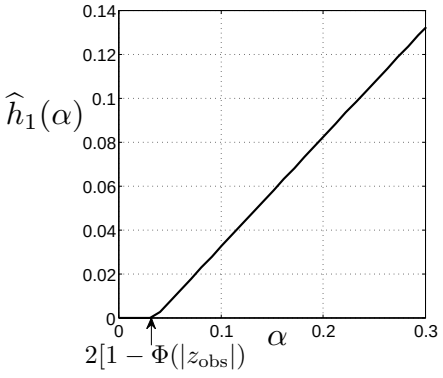

Figure 2: Robustness curve for problem 13a.

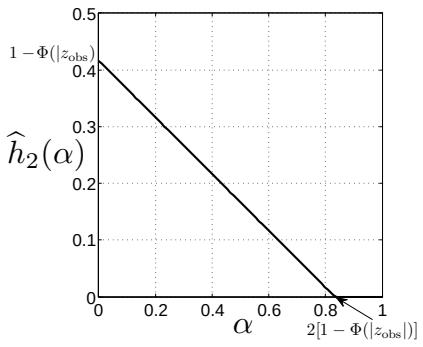

Figure 3: Robustness curve for problem 13b.

(13b) In analogy to eq.(104), the definition of the robustness is:

$$\widehat{h}_2(\alpha) = \max\left\{h:\ \left(\min_{F\in\mathcal{U}(h)} 2[1 - F(|z_{\rm obs}|)]\right) \geq \alpha\right\} \quad (107)$$

Let $m(h)$ denote the inner minimum, which occurs when $F(|z_{\rm obs}|)$ is as large as possible at horizon of uncertainty h:

$$F(|z_{\rm obs}|) = \min\left[1,\ \Phi(|z_{\rm obs}|) + h\right] \quad (108)$$

[3] Use matlab program \lectures\reltest\prob12.m to calculate robustness curves for both parts 13a and 13b

Thus, for $h \leq 1 - \Phi(|z_{\rm obs}|)$:

$$m(h) = 2\left[1 - \Phi(|z_{\rm obs}|) - h\right] \geq \alpha \quad \Longrightarrow \quad \boxed{\widehat{h}_2(\alpha) = \frac{2[1 - \Phi(|z_{\rm obs}|)] - \alpha}{2}} \tag{109}$$

or zero if this is negative. Note that, for all $0 \leq \alpha \leq 1$, $\widehat{h}$ does not exceed $1 - \Phi(|z_{\rm obs}|)$ so we need not evaluate $m(h)$ for larger values of h.

The robustness curve in eq.(109) is plotted in fig. 2 for the following parameter values: $\mu = 1$, $\sigma = 1.3$ and $N = 30$. The observed sample mean is $\overline{x}_{\rm obs} = 1.05$ for which $z_{\rm obs} = 0.2107$ and $\Phi(|z_{\rm obs}|) = 0.5834$.