Lecture Notes on Acceptance Testing

Yakov Ben-Haim
Yitzhak Moda'i Chair in Technology and Economics
Faculty of Mechanical Engineering
Technion — Israel Institute of Technology
Haifa 32000 Israel
yakov@technion.ac.il
http://info-gap.com   http://www.technion.ac.il/yakov

Primary source material: William W. Hines, and Douglas C. Montgomery, *Probability and Statistics in Engineering and Management Science.* 3rd ed. Sections 11-1–11-4, 11-11, 11-12.

Notes to the Student:

• These lecture notes are not a substitute for the thorough study of books. These notes are no more than an aid in following the lectures.

• Section 11 contains review exercises that will assist the student to master the material in the lecture and are highly recommended for review and self-study. The student is directed to the review exercises at selected places in the notes. They are not homework problems, and they do not entitle the student to extra credit.

Contents

1 Thickness Measurement: Testing a Sample Mean 3

2 Sequential Sampling: Testing a Mean 9

3 Matching Two Dimensions: Testing Two Sample Means 10

4 Engine Warming: A χ^2 Test 13

5 Failure Rate: Poisson Distribution and the χ^2 Test 16

6 χ^2 Test of Independence in a 2-way Table 18

7 Acceptance Sampling of a Large Population 22

8 Sample Size for Detecting a Change: Threshold Tests 26
8.1 Sample Size and Error Probabilities 26
8.2 Uncertain Effect Size and Variance 30
8.3 Uncertain Sample PDF 33
8.3.1 Background 33
8.3.2 Info-gap Approach to Determining the Sample Size 34
8.3.3 An Approximate Robustness for Small Effect Size 35

9 Tests of the Mean with Distributional Uncertainty 36
9.1 Distributional Uncertainty 36
9.2 Info-Gap Representations of Distributional Uncertainty 38
9.3 Robustness Functions with CDF Uncertainty 39
9.3.1 Binary Test: Formulation 39

[0] \lectures\reltest\acctes.tex 2.6.2016

9.3.2 Robustnesses for Type I Errors 40
9.3.3 Robustnesses for Type II Errors 42
9.3.4 Decisions and Judgments 43
9.4 Robustness Functions with PDF Uncertainty: Definitions 46
9.4.1 Binary Test: Formulation 46
9.4.2 Robustness for Falsely Rejecting H_0. (Type I Error.) 47
9.4.3 Robustness for Falsely Accepting H_0. (Type II Error.) 48
9.5 Robustness with PDF Uncertainty: Numerical Examples 49
9.5.1 Robustness for Type-I Error 49
9.5.2 Robustness for Type-II Error 49

10 Accelerated Lifetime Testing: Simple Case 51
10.1 Formulation 51
10.2 Uncertainty and Robustness 51
10.3 Evaluating the Robustness 52

11 Review Exercises 54

1 Thickness Measurement: Testing a Sample Mean

¶ In the first few sections we illustrate various **statistical hypothesis tests** for acceptance testing.

¶ Suppose we measure the thickness of a plate at N widely separated points: $x_1, \ldots, x_N$.

These measurements differ from one another due to random measurement error, as well as fluctuations in the local thickness.

How do we use these measurements to decide if the plate "really" has thickness T?

What does "really has thickness T" mean? Perhaps: $\mathrm{E}(T) = \mu$.

¶ A **random sample** is a set of independent measurements made on the same population. That is, a random sample is a set of independent and identically distributed (i.i.d.) random variables.

¶ The **sample mean** of a random sample is defined as:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1}$$

Theorem 1. If a random sample of size N is taken from a population with mean μ and variance σ^2, then the sample mean $\overline{x}$ has mean μ and variance σ^2/N. That is:

$$\mathrm{E}(\overline{x}) = \mu \tag{2}$$

$$\text{var}(\overline{x}) = \mathrm{E}\left([\overline{x} - \mathrm{E}(\overline{x})]^2\right) = \frac{\sigma^2}{N} \tag{3}$$

Proof. First consider eq.(2):

$$\mathrm{E}(\overline{x}) = \mathrm{E}\left(\frac{1}{N}\sum_{i=1}^{N} x_i\right) = \int_{x_1\ldots x_N} p(x_1, \ldots, x_N)\frac{1}{N}\sum_{i=1}^{N} x_i \,\mathrm{d}x_1\cdots x_N \tag{4}$$

Because the measurements of the random sample are independent:

$$p(x_1, \ldots, x_N) = \prod_{i=1}^{N} p(x_i) \tag{5}$$

Thus:

$$\mathrm{E}(\overline{x}) = \frac{1}{N}\sum_{i=1}^{N}\int p(x_i)x_i \,\mathrm{d}x_i = \frac{1}{N}\sum_{i=1}^{N}\mu = \mu \tag{6}$$

Which completes the proof of eq.(2). Note that this is independent of $p(x_i)$.

Now consider eq.(3):

$$\begin{aligned} \mathrm{var}(\overline{x}) &= \mathrm{E}[(\overline{x}-\mu)^2] = \mathrm{E}\left[\left(\frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)\right)^2\right] && (7) \\ &= \frac{1}{N^2}\sum_i\sum_j \mathrm{E}\left[(x_i-\mu)(x_j-\mu)\right] && (8) \\ &= \frac{1}{N^2}\sum_i \mathrm{E}(x_i-\mu)^2 = \frac{\sigma^2}{N} && (9) \end{aligned}$$

Note that this is independent of $p(x_i)$.

Review exercise 1, p. 54.

Review exercise 2, p. 54.

Theorem 2. If a random sample is taken from a normal population, then the sample mean is normal. ■

Combining the last two theorems we can assert:

$$x_i \sim \mathcal{N}(\mu, \sigma^2) \quad \Longrightarrow \quad \overline{x} \sim \mathcal{N}(\mu, \sigma^2/N) \tag{10}$$

Review exercise 3, p. 54.

"Theorem" 3. An approximate statement of the **central limit theorem:** The sample mean of a random sample will be approximately normal for large sample size ($N > 30$, rough number). ■

Review exercise 4, p. 54.

¶ So, let us suppose that the thickness measurement is normally distributed, or that the number of measurements is large. Thus:

$$\overline{x} \sim \mathcal{N}(\mu, \sigma^2/N) \tag{11}$$

where:
μ = true mean thickness.
σ^2 = variance of the thickness measurements.
N = sample size.

Also, assume that we **know the value of** σ^2.

¶ How do we decide whether:
- ○ The true thickness of the plate is T?
- ○ To accept or reject the plate?

We use an **hypothesis test.**

¶ The **null hypothesis:**

$$H_0 : \quad \mu = T \tag{12}$$

T = desired thickness: a known value.
μ = true thickness: an unknown value.

¶ The **alternative hypothesis:**

$$H_1 : \quad \mu \neq T \tag{13}$$

This is a **two-tailed test.** The test would be **one-tailed** if the alternative hypothesis were:

$$H_1 : \quad \mu > T \tag{14}$$

Or:

$$H_1 : \quad \mu < T \tag{15}$$

Review exercise 5, p. 54.

¶ **Level of confidence,** α**:**

- Probability of obtaining a result at least as extreme as the observed result, conditioned upon H_0.
- Probability of rejecting H_0 erroneously.

For the two-tailed test in question (see fig. 1 on p.5):

$$\alpha = \text{Prob}\left(|\overline{x} - T| \geq |\overline{x}_o - T| \,\middle|\, H_0 \right) \tag{16}$$

$\overline{x}$ = the random variable “sample mean”.
$\overline{x}_o$ = the observed value of the random variable $\overline{x}$.

Review exercise 6, p. 54.

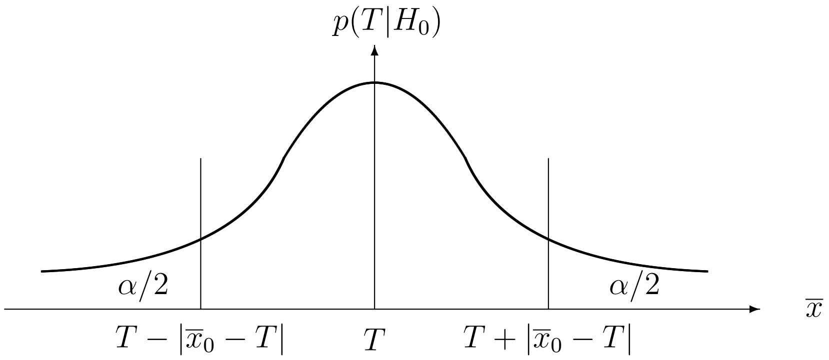


Figure 1: Sketch of probability density, illustrating the level of confidence in eq.(16).

¶ Interpreting the level of confidence:
If α is small: reject H_0.
If α is large: accept H_0.

¶ How to evaluate the level of confidence?

Conditioned upon H_0, we can assert:

$$\overline{x} \sim \mathcal{N}(T, \sigma^2/N) \tag{17}$$

$\overline{x}$ can be standardized as:

$$z = \frac{\overline{x} - T}{\sigma/\sqrt{N}} \sim \mathcal{N}(0,1) \tag{18}$$

Review exercise 7, p. 54.
Now the level of confidence can be expressed as (see fig. 2 on p.6):

$$\alpha = \text{Prob}\left(|\overline{x} - T| \geq |\overline{x}_o - T| \,\middle|\, H_0 \right) \tag{19}$$

$$= \text{Prob}\left(\frac{|\overline{x} - T|}{\sigma/\sqrt{N}} \geq \frac{|\overline{x}_o - T|}{\sigma/\sqrt{N}} \,\middle|\, H_0 \right) \tag{20}$$

$$= 2\left[1 - \Phi\left(\frac{|\overline{x}_o - T|}{\sigma/\sqrt{N}} \right)\right] \tag{21}$$

where $\Phi(\cdot)$ is the cdf of the standard normal distribution. Define:

$$z_o = \frac{|\overline{x}_o - T|}{\sigma/\sqrt{N}} \tag{22}$$

$$\alpha = 2[1 - \Phi(z_o)] \tag{23}$$

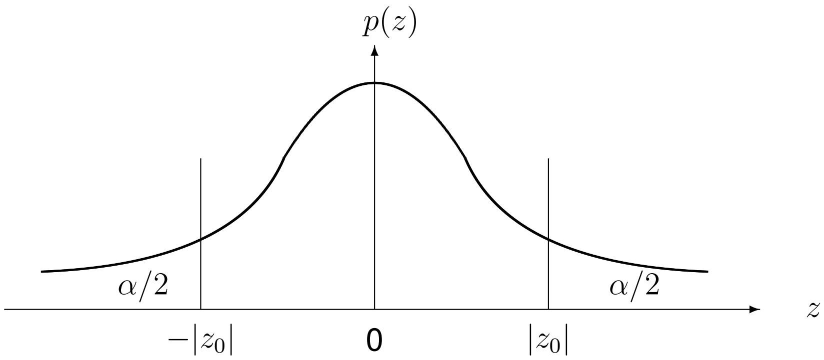


Figure 2: Sketch of probability density illustrating level of confidence in eq.(21).

¶ Numerical example.
Measurements: 1.3, 1.2, 1.4, 1.3, 1.1
Desired thickness: $T = 1.36$.
Known variance: $\sigma^2 = 0.01 \implies \sigma = 0.1$.

Hence: $N = 5$ and $\overline{x}_o = 1.26$.

$$z_o = \frac{1.26 - 1.36}{0.1/\sqrt{5}} = -2.236 \tag{24}$$

$$\alpha = \text{Prob}\left(|z| \geq |z_o| \,\middle|\, H_0\right) = 2\left[1 - \Phi(|z_o|)\right] = 0.024 \tag{25}$$

• Note: $\Phi(|z_o|) = 0.988$.

• So, the probability of getting a value as large or larger than the observed standardized sample mean is 0.024.

• This is rather small, so we tend to reject H_0.
If so, then we reject H_0 at the 0.024 level of confidence.

• Similarly, $0.024 =$ probability of falsely rejecting H_0.

¶ We tested H_0 on p. 5 under the assumption that σ^2 is known.

• What do we do if σ^2 is **unknown?**

• We will assume that the **sample is normal.**

¶ Two cases:
N large.
N small.

¶ Case 1: N large.
If $N \geq 25$ (rough number), then s^2, the sample variance, is a good estimate of the true population variance.

$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2 \tag{26}$$

We can now assume, conditioned on H_0, that:

$$t = \frac{\overline{x} - T}{s/\sqrt{N}} \sim \mathcal{N}(0, 1) \tag{27}$$

This assumption would be precise if $s^2 = \sigma^2$.
Now we proceed with the test as before.

Review exercise 8, p. 54.

¶ Case 2: N small.
If $N < 25$ (rough number), then the sample variance s^2 is not a good estimate of the population variance.
The statistic:

$$t = \frac{\overline{x} - T}{s/\sqrt{N}} \tag{28}$$

is broader than $\mathcal{N}(0, 1)$ since both $\overline{x}$ and s display variation.

This is a **t statistic** with $N-1$ degrees of freedom (dofs).
Recall we assumed the **sample is normal.**

¶ We repeat the numerical example, without knowledge of σ^2.
Measurements: 1.3, 1.2, 1.4, 1.3, 1.1
Desired thickness: $T = 1.36$.
$N = 5$ and $\overline{x}_o = 1.26$.
Sample variance: $s^2 = 0.013 \quad \Longrightarrow \quad s = 0.1140$.

The observed t statistic is:

$$t_o = \frac{\overline{x} - T}{s/\sqrt{N}} = \frac{1.26 - 1.36}{0.1140/\sqrt{5}} = -1.961 \tag{29}$$

The dofs: $5 - 1 = 4$.
There are many stats tables on the web. Mostly they are reliable, though I once found an erroneous table. This table seems fine:
`http://www.stats.gla.ac.uk/~levers/software/tables/tables-uog.pdf`
From a different table of the t distribution (transparency AS-p.7.1):

$\alpha =$	0.1	0.05	0.025	0.01
$\nu = 4:$	1.533	2.132	2.776	3.747

So, with 4 dofs:
The probability of t_4 exceeding 1.533 is 0.1.
The probability of t_4 exceeding 2.132 is 0.05.
Etc.

The level of confidence of this two-tailed test, with $t_o = -1.961$, is:

$$\alpha = \text{Prob}\left(|t| \geq |t_o| \,\middle|\, H_0 \right) \approx 2 \times 0.07 = 0.14 \tag{30}$$

This is not small, so we cannot reject H_0.
The probability of falsely rejecting H_0 is 0.14. ■
Review exercise 9, p. 54.

2 Sequential Sampling: Testing a Mean

¶ In the previous example we found that, with 5 measurements, we reject H_0 at 0.14 level of confidence.
This rejection is not very convincing. **Review exercise 10, p. 54.**
If we measured more, we could probably make a better, more confident decision.
How many measurements to make?
One approach is the idea of **sequential sampling:**
Continue adding measurements until the level of confidence is clear cut.
The following table shows an example.

N	x_i	$\overline{x}$	s^2	$s/\sqrt{N}$	$\|t_o\|$	α
5	1.3, 1.2, 1.4, 1.3, 1.1	1.26	0.013	0.0510	1.961	$2 \times 0.07 = 0.14$
8	1.1, 1.3, 1.2	1.2375	0.01125	0.0375	3.267	$2 \times 0.007 = 0.014$
11	1.1, 1.2, 1.1	1.2091	0.01091	0.0315	4.7915	< 0.002

Table 1: Data from a sequential test of the mean thickness measurement. (Transparency)

After 11 measurements we can stop:
Our confidence in rejecting H_0 is great.

Review exercise 11, p. 54.

¶ General theory: sequential analysis.[1]

[1] Abraham Wald, *Sequential Analysis.*

3 Matching Two Dimensions: Testing Two Sample Means

¶ Let us suppose that we are measuring two dimensions, to see if they match. For instance, the inner and outer dimensions of pieces that need to fit together.

These two dimensions are measured repeatedly, with a measuring device which has random errors:
- ○ The random sample of the inner dimension is: $x_1, \ldots, x_N$.
- ○ The random sample of the outer dimension is: $y_1, \ldots, y_M$.

We need not assume that the sample sizes, N and M are the same.

The true mean values of these two samples, which are the true dimensions, are:
μ_1 = true (but unknown) inner dimension.
μ_2 = true (but unknown) outer dimension.

¶ We assume that these two samples each have the same **known variance** σ^2.

Will the pieces fit snugly?

How confident are we of the answer?
In other words, we wish to test the null hypothesis:

$$H_0 : \quad \mu_1 = \mu_2 \tag{31}$$

against one of the following alternative hypotheses:

$$H_1 : \quad \mu_1 \neq \mu_2 \tag{32}$$

or:

$$H_1 : \quad \mu_1 > \mu_2 \tag{33}$$

or:

$$H_1 : \quad \mu_1 < \mu_2 \tag{34}$$

We choose an alternative hypothesis depending upon our prior information.

Review exercise 12, p. 54.

Let $\overline{x}$ and $\overline{y}$ be the sample means:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \overline{y} = \frac{1}{M}\sum_{i=1}^{M} y_i \tag{35}$$

Consider the statistic:

$$\Delta = \overline{x} - \overline{y} \tag{36}$$

What is the mean and variance of Δ, if H_0 holds? If H_1 holds?
Note that we cannot answer this question under H_1.

Review exercise 13, p. 54.

$$\mathrm{E}(\Delta) = \mathrm{E}(\overline{x} - \overline{y}) = \mathrm{E}(\overline{x}) - \mathrm{E}(\overline{y}) = 0 \tag{37}$$

$$\mathrm{var}(\Delta) = \mathrm{var}(\overline{x} - \overline{y}) = \mathrm{var}(\overline{x}) + \mathrm{var}(\overline{y}) = \underbrace{\frac{\sigma^2}{N} + \frac{\sigma^2}{M}}_{\sigma_\Delta^2} \tag{38}$$

In eq.(38) we have used the fact that $\overline{x}$ and $\overline{y}$ are statistically independent because they are means of separate random samples:

$$\begin{aligned}
\mathrm{var}(\overline{x} - \overline{y}) &= \mathrm{E}\left[\left(\overline{x} - \overline{y} - [\mathrm{E}(\overline{x} - \overline{y})]\right)^2\right] && (39)\\
&= \mathrm{E}\left[\left(\overline{x} - \mathrm{E}(\overline{x})\right)^2\right] - 2\mathrm{E}\left[\left(\overline{x} - \mathrm{E}(\overline{x})\right)\left(\overline{y} - \mathrm{E}(\overline{y})\right)\right] + \mathrm{E}\left[\left(\overline{y} - \mathrm{E}(\overline{y})\right)^2\right] && (40)\\
&= \mathrm{E}\left[\left(\overline{x} - \mathrm{E}(\overline{x})\right)^2\right] + \mathrm{E}\left[\left(\overline{y} - \mathrm{E}(\overline{y})\right)^2\right] && (41)\\
&= \mathrm{var}(\overline{x}) + \mathrm{var}(\overline{y}) && (42)
\end{aligned}$$

How is Δ distributed if H_0 is true?
If we assume either of the following:
- ◦ The samples are large, **or**
- ◦ The measurements are normally distributed.

Then, in either case:

$$\Delta \sim \mathcal{N}(0, \sigma_\Delta^2) \tag{43}$$

or equivalently:

$$\frac{\Delta}{\sigma_\Delta} \sim \mathcal{N}(0, 1) \tag{44}$$

Review exercise 14, p. 54.

Assume **either** normality **or** large samples, and define:

$$z = \frac{\Delta}{\sigma_\Delta} \tag{45}$$

If H_0 holds, then:

$$z \sim \mathcal{N}(0, 1) \tag{46}$$

Review exercise 15, p. 54.

Using the two-sided alternative hypothesis of eq.(32) on p.10, we formulate the level of confidence as (fig. 3, p.12):

$$\alpha = \text{Prob}\left(|z| \geq |z_o| \,\middle|\, H_0\right) \tag{47}$$

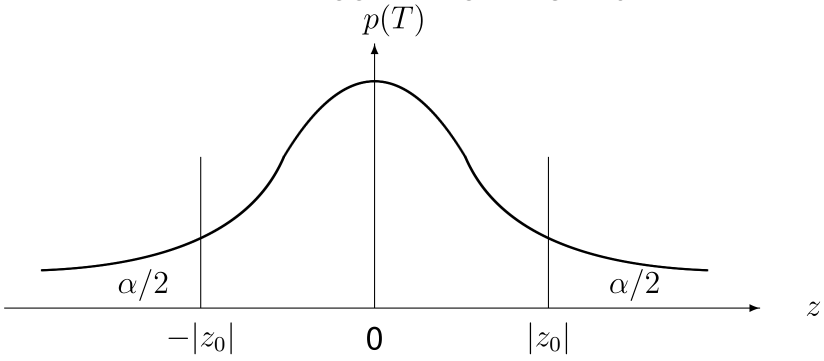


Figure 3: Sketch of probability density illustrating level of confidence in eq.(47).

Or, if we use the 1-sided alternative hypothesis of eq.(33), the level of confidence becomes (fig. 4, p.12):

$$\alpha = \text{Prob}\left(z \geq z_o \,\middle|\, H_0\right) \tag{48}$$

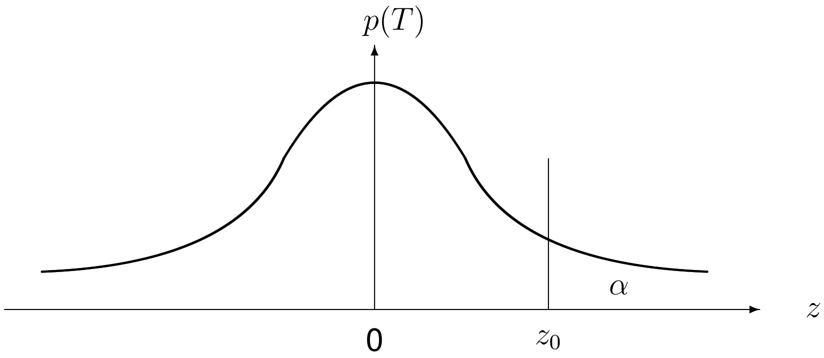


Figure 4: Sketch of probability density illustrating level of confidence in eq.(48).

In the 2-tailed case we obtain:

$$\alpha = 2\left[1 - \Phi(z_o)\right] \tag{49}$$

In the 1-tailed case we obtain:

$$\alpha = 1 - \Phi(z_o) \tag{50}$$

4 Engine Warming: A χ^2 Test

¶ We use an example to introduce the idea of a χ^2 hypothesis test.

¶ The temperature of an operating engine fluctuates between "warm" and "hot". For normal operation the engine should be "hot" a fraction $p_{\rm h} = 0.15$ of the time.

Maintenance records since the last overhaul of the engine show that the engine was "warm" at $N_{\rm w} = 162$ and "hot" at $N_{\rm h} = 44$ statistically independent sample instants.

$$\frac{N_{\rm h}}{N_{\rm h} + N_{\rm w}} = \frac{44}{206} \approx 0.21 \tag{51}$$

Is the engine operating properly?

Review exercise 16, p. 54.

¶ The χ^2 **test for categorical data** is suitable for addressing this question.

We formulate the χ^2 test as follows.
Given:
- ○ K types of outcomes of an 'experiment.'
- ○ n_i outcomes of type i, $i = 1, \ldots, K$.
- ○ $N = \sum_{i=1}^{K} n_i$ = total number of outcomes.

Null Hypothesis:

$$H_0: \quad p_i = \text{probability of type } i \text{ outcome}, i = 1, \ldots, K \tag{52}$$

where $p_1, \ldots, p_K$ are known values.
The alternative hypothesis, H_1, is simply: H_0 is false. That is, at least one of the probabilities in H_0 is wrong.
The χ^2 statistic is:

$$\chi^2 = \sum_{i=1}^{K} \frac{(n_i - Np_i)^2}{Np_i} \tag{53}$$

Explanation:
- ○ The numerator is a prediction-error for type-i outcomes.
- ○ We expect χ^2 to be small if H_0 is correct.

Theorem 4: If N is large and if H_0 is true, then the statistic in eq.(53) is approximately a χ^2 random variable with $K - 1$ dofs. ■

¶ The χ^2 distribution is shown in fig. 5, p.14, for several values of the dof.

¶ How does one calculate the level of confidence, α?

- Recall:
  - ○ α is the probability, conditioned upon the null hypothesis, of obtaining a value more

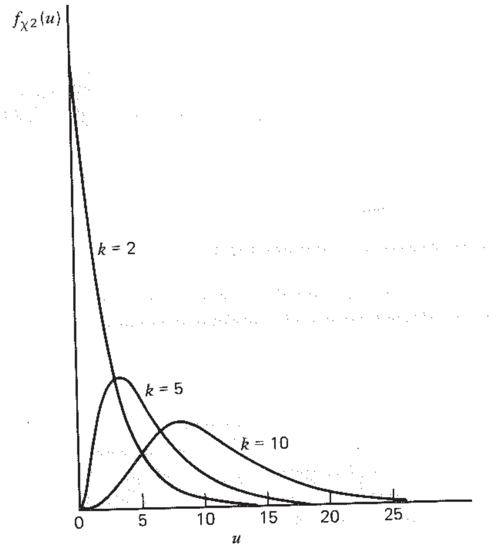


Figure 9-1 Several χ^2 distributions.

Figure 5: Chi squared distributions. Hines and Montgomery, p.232.

extreme than the observed statistic.

∘ $\alpha =$ probability of falsely rejecting H_0.

Hence, let χ_o^2 be the observed value of the χ^2 statistic of eq.(53). The level of confidence is:

$$\alpha = \text{Prob}\left(\chi^2 \geq \chi_o^2 \,\middle|\, H_0\right) \tag{54}$$

α = 'small' $\Longrightarrow$ χ_o^2 = 'extreme' $\Longrightarrow$ Reject H_0.
α = 'large' $\Longrightarrow$ χ_o^2 = 'not extreme' $\Longrightarrow$ Accept H_0.

'Small' and 'Large' are understood from the natural calibration of probability: from 0 to 1.

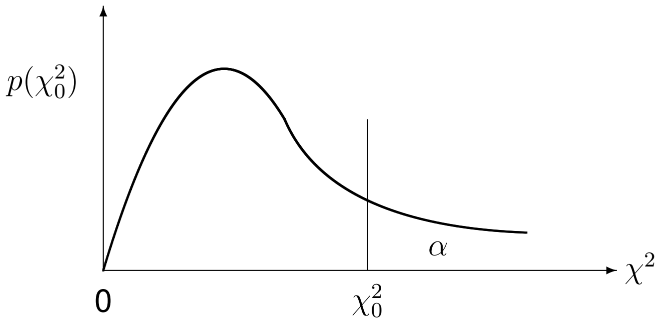


Figure 6: Level of confidence in eq.(54)

¶ Let us apply this theorem to our example.
$K = 2$: 2 possible states: 'hot' and 'warm'.
$N_\text{w} = 162, \quad N_\text{h} = 44, \quad N = 206.$
H_0: $p_\text{h} = 0.15, \ p_\text{w} = 0.85.$

$$\chi_o^2 = \frac{(162 - 206 \times 0.85)^2}{206 \times 0.85} + \frac{(44 - 206 \times 0.15)^2}{206 \times 0.15} = 6.533 \quad (55)$$

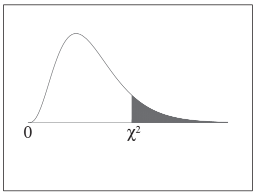
DOFs $= 2 - 1 = 1$.
χ^2 table (see the table on p.15):

$\alpha =$	0.05	0.025	0.01
$\nu = 1:$	3.84	5.02	6.63

From this we see that $\alpha \approx 0.01$.
So, we reject H_0 at the 1% confidence level.
In other words, the evidence is strong that the engine is running 'hot'.

Review exercise 17, p. 54.

Chi-Square Distribution Table



The shaded area is equal to α for $\chi^2 = \chi^2_\alpha$.

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750
6	0.676	0.872	1.237	1.635	2.204	10.645	12.592	14.449	16.812	18.548
7	0.989	1.239	1.690	2.167	2.833	12.017	14.067	16.013	18.475	20.278
8	1.344	1.646	2.180	2.733	3.490	13.362	15.507	17.535	20.090	21.955
9	1.735	2.088	2.700	3.325	4.168	14.684	16.919	19.023	21.666	23.589
10	2.156	2.558	3.247	3.940	4.865	15.987	18.307	20.483	23.209	25.188
11	2.603	3.053	3.816	4.575	5.578	17.275	19.675	21.920	24.725	26.757
12	3.074	3.571	4.404	5.226	6.304	18.549	21.026	23.337	26.217	28.300
13	3.565	4.107	5.009	5.892	7.042	19.812	22.362	24.736	27.688	29.819
14	4.075	4.660	5.629	6.571	7.790	21.064	23.685	26.119	29.141	31.319
15	4.601	5.229	6.262	7.261	8.547	22.307	24.996	27.488	30.578	32.801
16	5.142	5.812	6.908	7.962	9.312	23.542	26.296	28.845	32.000	34.267
17	5.697	6.408	7.564	8.672	10.085	24.769	27.587	30.191	33.409	35.718
18	6.265	7.015	8.231	9.390	10.865	25.989	28.869	31.526	34.805	37.156
19	6.844	7.633	8.907	10.117	11.651	27.204	30.144	32.852	36.191	38.582
20	7.434	8.260	9.591	10.851	12.443	28.412	31.410	34.170	37.566	39.997
21	8.034	8.897	10.283	11.591	13.240	29.615	32.671	35.479	38.932	41.401
22	8.643	9.542	10.982	12.338	14.041	30.813	33.924	36.781	40.289	42.796
23	9.260	10.196	11.689	13.091	14.848	32.007	35.172	38.076	41.638	44.181
24	9.886	10.856	12.401	13.848	15.659	33.196	36.415	39.364	42.980	45.559
25	10.520	11.524	13.120	14.611	16.473	34.382	37.652	40.646	44.314	46.928
26	11.160	12.198	13.844	15.379	17.292	35.563	38.885	41.923	45.642	48.290
27	11.808	12.879	14.573	16.151	18.114	36.741	40.113	43.195	46.963	49.645
28	12.461	13.565	15.308	16.928	18.939	37.916	41.337	44.461	48.278	50.993
29	13.121	14.256	16.047	17.708	19.768	39.087	42.557	45.722	49.588	52.336
30	13.787	14.953	16.791	18.493	20.599	40.256	43.773	46.979	50.892	53.672
40	20.707	22.164	24.433	26.509	29.051	51.805	55.758	59.342	63.691	66.766
50	27.991	29.707	32.357	34.764	37.689	63.167	67.505	71.420	76.154	79.490
60	35.534	37.485	40.482	43.188	46.459	74.397	79.082	83.298	88.379	91.952
70	43.275	45.442	48.758	51.739	55.329	85.527	90.531	95.023	100.425	104.215
80	51.172	53.540	57.153	60.391	64.278	96.578	101.879	106.629	112.329	116.321
90	59.196	61.754	65.647	69.126	73.291	107.565	113.145	118.136	124.116	128.299
100	67.328	70.065	74.222	77.929	82.358	118.498	124.342	129.561	135.807	140.169

Figure 7: Chi squared table.

5 Failure Rate: Poisson Distribution and the χ^2 Test

¶ A computer controlled milling machine operates automatically except when jamming, tool breakage or other failures occur. Under normal circumstances these failures occur at a rate of about 1 or 2 per day. Also, the distribution in time of the failures is **thought to be** a Poisson process: (1) constant average failure rate; (2) events occur independently.
¶ Data have accumulated over a 50-day period for this machine. In this time 75 failures occurred, so the average failure rate is:

$$\lambda = \frac{75 \text{ failures}}{50 \text{ days}} = 1.5 \frac{\text{failures}}{\text{day}} \tag{56}$$

Furthermore, we know how many days had 0, 1, 2 and 3 or more failures:

Failures/day, i	# days, n_i	# failures, p_i
0	12	0
1	17	17
2	9	18
3+	12	40
Totals:	50	75

Table 2: Failure data.

¶ We want to test the hypothesis: the distribution over time of failures is described by a Poisson process.

Recall the Poisson distribution:
P_i = probability of exactly i failures in a single day.

$$P_i = \frac{e^{-\lambda}\lambda^i}{i!}, \quad i = 0, 1, 2, \ldots \tag{57}$$

¶ We can use a χ^2 test to test this hypothesis.
First define some notation:
N = total number of days.
$p_i = P_i$, the Poisson distribution in H_0, for $i = 0, \ldots, 2$.
$p_3 = \sum_{i=3}^{\infty} P_i$.
n_i = observed number of days with i failures, $i = 0, \ldots, 3$.

Np_i = expected number of days with i failures, $i = 0, \ldots, 3$.

Now the null hypothesis is:
H_0: distribution of failures is p_i with $\lambda = 1.5$/day.
H_1: H_0 is wrong.

¶ The χ^2 statistic is:

$$\chi_o^2 = \sum_{i=0}^{3} \frac{(n_i - Np_i)^2}{Np_i} = 1.695 \tag{58}$$

The DOFs:

$$DOF = \underbrace{4}_{\text{catagories}} - \underbrace{1}_{\text{normalization}} - \underbrace{1}_{\text{estimating } \lambda} = 2 \tag{59}$$

¶ Recall the pth quantile with ν DOFs, $\chi^2_{(\nu),p}$:

$$p = \text{Prob}\left(\chi^2_{(\nu)} \le \chi^2_{(\nu),p}\right) \tag{60}$$

From a χ^2 table, the p-quantiles for $\nu = 2$ are:

$p =$	0.5	0.6
$\chi^2_{(2),p}$ =	1.386	1.833

From this table we see that the level of confidence, with $\chi^2_o = 1.695$, is:

$$\alpha = \text{Prob}\left(\chi^2 \ge \chi^2_o \,\middle|\, H_0\right) \approx 1 - 0.55 = 0.45 \tag{61}$$

This is very large, so we accept H_0: the distribution of failures is Poisson.

¶ Now consider a **different machine,** for which the data are:

Failures/day, i	# days, n_i	# failures, p_i
0	15	0
1	13	13
2	8	16
3+	14	46
Totals:	50	75

Table 3: Failure data.

As before, the average failure rate is:

$$\lambda = 1.5 \frac{\text{failures}}{\text{day}} \tag{62}$$

With these results the observed χ^2 value is:

$$\chi^2 = 5.86 \tag{63}$$

With 2 DOFs, this implies that the level of confidence is:

$$\alpha = \text{Prob}\left(\chi^2 \ge \chi^2_o \,\middle|\, H_0\right) \approx 0.06 \tag{64}$$

This is rather small, so we reject H_0: the distribution of failures is not Poisson.

Some factor is causing either:
- ○ Variation in the average failure rate.
- ○ Failure inter-dependence.

In short: clustering of failures.

6 χ^2 Test of Independence in a 2-way Table

¶ Consider the following situation:
Two different systems are used to perform a particular mission. Records indicate the number of failed and successful missions:

	Successful Missions	Failed Missions	Total Missions
System 1	86	35	121
System 2	64	37	101
Total Missions	150	72	

Table 4: Data on successes and failures of two systems.

¶ **Question:**
Is there solid evidence for the contention that system 1 is more reliable than system 2?

In other words, are the rows and columns of the table dependent?
That is, by choosing the row (system), do I have sound evidence that I significantly influence the column (success or failure) in which I "fall"?

We can use the χ^2 test to evaluate the evidence.
¶ First we must formulate the **null hypothesis.**
p_{ij} = probability of an event in row i, column j.
$p_{i\bullet}$ = probability of an event in row i.
$p_{\bullet j}$ = probability of an event in column j.

The null hypothesis states:
There is statistical independence between rows and columns.

That is:

$$H_0 : \quad p_{ij} = p_{i\bullet} p_{\bullet j} \tag{65}$$

The alternative hypothesis states that H_0 is false.

¶ Now we can formulate the χ^2 statistic. Define:
n_{ij} = number of events in row i and column j.
r = number of rows.
c = number of columns.
N = total number of events = $\sum_{i=1}^{r} \sum_{j=1}^{c} n_{ij}$.
The χ^2 statistic is calculated as:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(n_{ij} - N p_{i\bullet} p_{\bullet j})^2}{N p_{i\bullet} p_{\bullet j}} \tag{66}$$

The level of confidence is the probability of χ^2 obtaining a value greater than the observed value, χ_o^2, conditioned on H_0:

$$\alpha = \text{Prob}\left(\chi^2 \geq \chi_o^2 \,\middle|\, H_0\right) \tag{67}$$

¶ **A difficulty:** in our example we don't know the values of $p_{i\bullet}$ and $p_{\bullet j}$, so we can't calculate χ_o^2.

Solution: estimate $p_{i\bullet}$ and $p_{\bullet j}$ from the data: n_{ij}:

$$\widehat{p}_{i\bullet} = \frac{n_{i\bullet}}{N}, \quad \widehat{p}_{\bullet j} = \frac{n_{\bullet j}}{N} \tag{68}$$

where:
$n_{i\bullet}$ = sum of ith row.
$n_{\bullet j}$ = sum of jth column.

¶ How many DOFs do we have?
DOF = number of categories $-$ number of constraints.
The number of categories is rc.
3 types of constraints:
Type 1. 1 constraint (normalization):

$$\sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij} = N \tag{69}$$

Type 2. $r-1$ constraints:
Estimate $\widehat{p}_{1\bullet}, \ldots, \widehat{p}_{r-1\bullet}$.
$\left(\widehat{p}_{r\bullet} = 1 - \sum_{i=1}^{r-1} \widehat{p}_{i\bullet}\right)$.

Type 3. $c-1$ constraints:
Estimate $\widehat{p}_{\bullet 1}, \ldots, \widehat{p}_{\bullet c-1}$.
$\left(\widehat{p}_{\bullet c} = 1 - \sum_{j=1}^{c-1} \widehat{p}_{\bullet j}\right)$.

So the number of DOFs is:

$$DOF = rc - 1 - (r-1) - (c-1) = (r-1)(c-1) \tag{70}$$

¶ Now the observed statistic can be calculated as:

$$\chi_o^2 = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - N\widehat{p}_{i\bullet}\widehat{p}_{\bullet j})^2}{N\widehat{p}_{i\bullet}\widehat{p}_{\bullet j}} \tag{71}$$

¶ In our numerical example we have:
$n_{11} = 86, \quad n_{12} = 35, \quad n_{1\bullet} = 121$

$n_{21} = 64, \quad n_{22} = 37, \quad n_{2\bullet} = 101$
$n_{\bullet 1} = 150, \quad n_{\bullet 2} = 72, \quad N = 222$

Hence the estimated probabilities are:

$$\widehat{p}_{1\bullet} = \frac{n_{1\bullet}}{N} = 0.545, \quad \widehat{p}_{2\bullet} = \frac{n_{2\bullet}}{N} = 0.455 \tag{72}$$

$$\widehat{p}_{\bullet 1} = \frac{n_{\bullet 1}}{N} = 0.676, \quad \widehat{p}_{\bullet 2} = \frac{n_{\bullet 2}}{N} = 0.324 \tag{73}$$

The observed statistic becomes:

$$\chi^2_o = 1.493 \tag{74}$$

And the number of DOFs is:

$$DOF = (2-1)(2-1) = 1 \tag{75}$$

Is 1.493 large or small? Should we accept or reject H_0?

We need to calibrate χ^2_o using the **quantile** values of the χ^2 distribution with 1 DOF.
Define:
$\chi^2_{(\nu)}$ = χ^2 random variable with ν DOFs.
$\chi^2_{(\nu),p}$ = pth quantile of $\chi^2_{(\nu)}$: fig. 8.
$\chi^2_{(\nu),p}$ is defined in:

$$p = \text{Prob}\left(\chi^2_{(\nu)} \le \chi^2_{(\nu),p}\right) \tag{76}$$

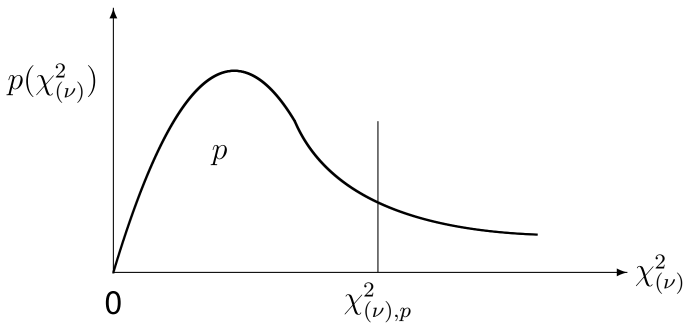


Figure 8: pth quantile of $\chi^2_{(\nu)}$, eq.(76)

From a χ^2 table, the p-quantiles for $\nu = 1$ are:

$p =$	0.75	0.80
$\chi^2_{(1),p}$ =	1.323	1.642

From this table we see that:

$$\text{Prob}\left(\chi^2_{(1)} \ge 1.323\right) = 1 - 0.75 = 0.25 \tag{77}$$

$$\text{Prob}\left(\chi^2_{(1)} \ge 1.642\right) = 1 - 0.80 = 0.20 \tag{78}$$

So, since $\chi^2_o = 1.493$ we see that:

$$\alpha \approx 0.22 \tag{79}$$

We cannot reject H_0 at 0.2 level of confidence.
So we accept H_0:
 ◦ Columns and rows are independent.
 ◦ The two systems are not significantly different in reliability.

Review exercise 18, p. 54.

¶ **Suppose we have many more observations, but in the same ratios** as table 4, p.18.

- See "big sample" data in table 5.

	Successful Missions	Failed Missions	Total Missions
System 1	8600	3500	12100
System 2	6400	3700	10100
Total Missions	15000	7200	

Table 5: Big-sample data on successes and failures of two systems.

- From eq.(68), p.19, we see that $\widehat{p}_{ij}$ **does not change.**

- However, $N_{\rm big} = 100 N_{\rm small}$ and $n_{ij,{\rm big}} = 100 n_{ij,{\rm small}}$.

- Thus, from eq.(66), p.18: $\chi^2_{\rm big,obs} = 100\chi^2_{\rm small,obs} = 149.3$ with same DOF.

- Hence: $\alpha_{\rm big,obs} \ll 0.001$ and we **now reject H_0 very strongly.**

7 Acceptance Sampling of a Large Population

¶ We have a large batch of items, among which an unknown fraction p are defective. We wish to sample this population to decide whether or not to accept the batch.

¶ Define:
N = sample size.
$p_{\rm a}$ = acceptable proportion of defective items.
$p_{\rm u}$ = unacceptable proportion of defective items.

$$p_{\rm u} > p_{\rm a} \tag{80}$$

$p_{\rm a}$ and $p_{\rm u}$ can be interpreted:
- ∘ We desire the defective fraction to be no greater than $p_{\rm a}$.
- ∘ We are willing to live with a defective fraction as large as $p_{\rm u}$.

Review exercise 19, p. 55.

p = true but unknown fraction of defective items in the batch.
$P_{\rm A}$ = probability of accepting the batch.

¶ Our algorithm for accepting or rejecting the batch is:
Accept if and only if:

$$\frac{\text{number of defective items in sample}}{\text{sample size}} \le p_{\rm a} \tag{81}$$

¶ Two types of errors can be made:
I. Acceptance with $p > p_{\rm u}$. Bad acceptance.
II. Rejection with $p < p_{\rm a}$. Bad rejection.

Review exercise 20, p. 55.

¶ How do we calculate $P_{\rm A}$, the probability of acceptance?

Assume that $N \ll$ batch size. Thus the sample does not significantly change the composition of the population.

¶ Binomial distribution:
$b(i;p,N)$ = probability of exactly i defectives in a sample of size N.

$$b(i;p,N) = \binom{N}{i} p^i (1-p)^{N-i}, \quad \binom{N}{i} = \frac{N!}{i!(N-i)!} \tag{82}$$

Review exercise 21, p. 55.

¶ Acceptance probability:

$$P_{\mathrm{A}} = \sum_{i=0}^{p_{\mathrm{a}}N} b(i;p,n) \quad (83)$$

$$= \sum_{i=0}^{p_{\mathrm{a}}N} \binom{N}{i} p^i (1-p)^{N-i} \quad (84)$$

¶ **Example.** Suppose:
$N = 100$.
p_{a} = 0.01
Hence:
$p_{\mathrm{a}}N = 1$

$$P_{\mathrm{A}} = \sum_{i=0}^{1} b(i;p,100) = (1-p)^{100} + 100p(1-p)^{99} \quad (85)$$

A plot of P_{A} vs. p reveals the probabilities of type I and type II errors.

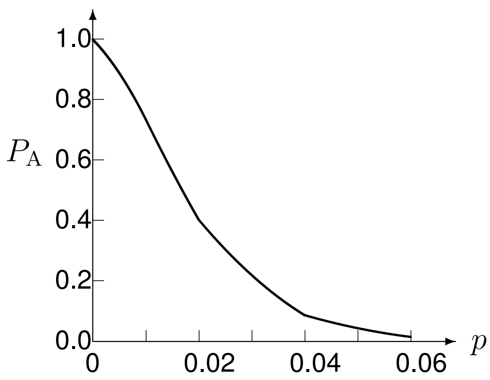


Figure 9: Acceptance probability in eq.(85)

Recall that:
p_{a} = max “acceptable” fraction of defectives.
p_{u} = max “tolerable” fraction of defectives.

¶ **Type I** errors:
Acceptance with $p > p_{\mathrm{u}}$. Bad acceptance.

The probability of a type I error is called the **consumer’s risk:** (fig. 10, p.24)

$$P_{\mathrm{I}} = P_{\mathrm{A}}(p = p_{\mathrm{u}}) \quad (86)$$

¶ **Type II** errors:
Rejection with $p < p_{\mathrm{a}}$. Bad rejection.

The probability of a type II error is called the **producer’s risk:** (fig. 10, p.24)

$$P_{\mathrm{II}} = 1 - P_{\mathrm{A}}(p = p_{\mathrm{a}}) \quad (87)$$

Review exercise 22, p. 55.

When we are designing a sampling scheme, we can evaluate it with 2 pairs of numbers:
$(p_u,\ P_I), \quad (p_a,\ P_{II})$

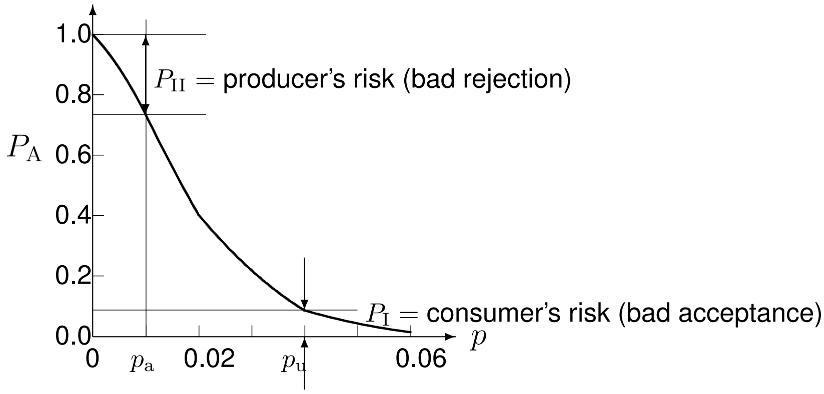


Figure 10: Illustration of type I and type II errors in eqs.(86) and (87) using eq.(85) ($N = 100$). $p_a = 0.01$, $p_u = 0.04$.

¶ **Example.** Consider $P_A(p)$ in eq.(85) on p.23.
Choose p_a = 0.01 and p_u = 0.04.

With $N =$ **100**, we have: $p_a N = 1$:
Consumer's risk: $P_I = P_A(p = p_u) =$ **0.0872**.
Producer's risk: $P_{II} = 1 - P_A(p = p_a) = 1 - 0.736 =$ **0.264**.

With $N =$ **200**, we have: $p_a N = 2$:

$$P_A = \sum_{i=0}^{2} b(i; p, 100) = (1-p)^{200} + 200p(1-p)^{199} + \frac{(200)(199)}{2} p^2 (1-p)^{198} \tag{88}$$

Consumer's risk: $P_I = P_A(p = p_u) =$ **0.0125**.
Producer's risk: $P_{II} = 1 - P_A(p = p_a) = 1 - 0.677 =$ **0.323**.

Review exercise 23, p. 55.

By **increasing** the sample size we:
- ○ **Increased** the producer's risk: 0.264 $\to$ 0.323.
- ○ **Decreased** the consumer's risk: 0.0872 $\to$ 0.0125.

¶ At $N \to \infty$ we expect rectangular P_A vs. p:
Consumer's risk: $P_I = P_A(p = p_u) = 0$.
Producer's risk: $P_{II} = 1 - P_A(p = p_a) = 0$.

For finite N, $P_A(p)$ oscillates with N, as a result of the discrete, binary nature of the distribution.

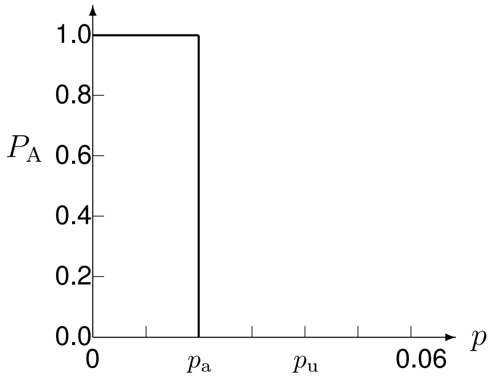


Figure 11: Asymptotical acceptance probability, $N \to \infty$.

Review exercise 24, p. 55.

8 Sample Size for Detecting a Change: Threshold Tests

Source material: David R. Fox, Yakov Ben-Haim, Keith R. Hayes, Michael McCarthy, Brendan Wintle, Piers Dunstan, 2007, An info-gap approach to power and sample size calculations, *Environmetrics,* vol. 18, pp.189–203.

8.1 Sample Size and Error Probabilities

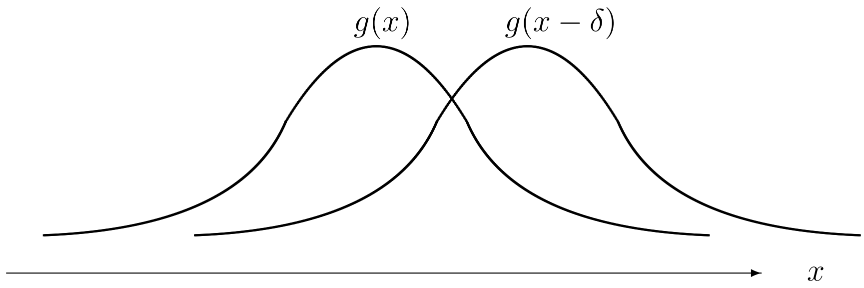


Figure 12: Pdf's $g(x)$ and $g(x-\delta)$ shifted to the right by δ.

¶ **The dispute.**
• x is the measurement of the output of a system.
• One side argues: x has not changed, its mean is μ_0 and its pdf is $g(x)$.
• The other side argues: x has changed, its mean is $\mu_1 = \mu_0 + \delta$ and its pdf has shifted to $g(x-\delta)$.
• $\delta > 0$ means that $g(x-\delta)$ is shifted to the right of $g(x)$.
• δ is the "effect size": the shift in the distribution of x.

Review exercise 25, p. 55.

• Null and alternative hypotheses:

$$H_0 : \qquad x \sim g(x) \tag{89}$$
$$H_1 : \qquad x \sim g(x-\delta) \tag{90}$$

¶ **The question:** How large a sample is needed to confidently resolve the dispute?

¶ **Random sample.**
• A random sample of x-values is taken. $n =$ sample size.
• $\overline{x}$ = sample mean. $f_n(\overline{x})$ is the pdf of the sample mean.
• $f_n(\overline{x})$ may differ from $g(x)$. E.g., large n implies $f_n(\overline{x})$ is normal regardless of $g(x)$.
• Null and alternative hypotheses:

$$H_0 : \qquad \overline{x} \sim f_n(\overline{x}) \tag{91}$$
$$H_1 : \qquad \overline{x} \sim f_n(\overline{x}-\delta) \tag{92}$$

• $\delta > 0$ means that $f_n(\overline{x}-\delta)$ is shifted δ to the right of $f_n(\overline{x})$.

¶ **Example:** Normal distribution:

$$x \sim \mathcal{N}(\mu, \sigma^2) \quad \text{implies} \quad \overline{x} \sim \mathcal{N}(\mu, \sigma^2/n) \tag{93}$$

Review exercise 26, p. 55.

- Hence the null and alternative hypotheses are:

$$H_0: \qquad \overline{x} \sim \mathcal{N}(\mu, \sigma^2/n) \tag{94}$$

$$H_1: \qquad \overline{x} \sim \mathcal{N}(\mu+\delta, \sigma^2/n) \tag{95}$$

- The pdf of the sample mean depends on the sample size.

¶ **Critical value,** C:
- Reject H_0 iff $\overline{x} > C$.
- C depends on the sample size.

¶ **Errors:**
- Type I error: false rejection of H_0.
- Type II error: false rejection of H_1.

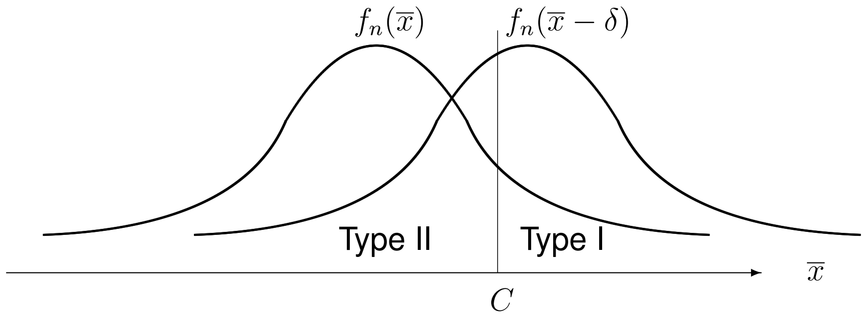


Figure 13: Pdf's for H_0 and H_1, illustrating type-I and type-II errors.

¶ **Error Probabilities:**
- $\alpha =$ probability of falsely rejecting H_0:

$$\alpha = \text{Prob}(\overline{x} > C | H_0) \tag{96}$$

$$= \int_C^\infty f_n(\overline{x})\, d\overline{x} \tag{97}$$

- $\beta =$ probability of falsely rejecting H_1 = probability of falsely accepting H_0:

$$\beta = \text{Prob}(\overline{x} \le C | H_1) \tag{98}$$

$$= \int_{-\infty}^{C} f_n(\overline{x}-\delta)\, d\overline{x} \tag{99}$$

$$= \int_{-\infty}^{C-\delta} f_n(\overline{x})\, d\overline{x} \tag{100}$$

- Power = $1-\beta =$ probability of correctly rejecting H_1.

¶ **Example: Normal distribution.**

- Null and alternative hypotheses are as in eqs.(94) and (95), p.27.
- Probability of type I error:

$$\alpha = \text{Prob}(\overline{x} > C|H_0) \tag{101}$$
$$= \text{Prob}\left(\frac{\overline{x}-\mu}{\sigma/\sqrt{n}} > \frac{C-\mu}{\sigma/\sqrt{n}}\right) \tag{102}$$
$$= 1 - \Phi\left(\frac{C-\mu}{\sigma/\sqrt{n}}\right) \tag{103}$$

where Φ is the cumulative distribution function of the standard normal distribution.

- Probability of type II error:

$$\beta = \text{Prob}(\overline{x} \le C|H_1) \tag{104}$$
$$= \text{Prob}\left(\frac{\overline{x}-(\mu+\delta)}{\sigma/\sqrt{n}} \le \frac{C-(\mu+\delta)}{\sigma/\sqrt{n}}\right) \tag{105}$$
$$= \Phi\left(\frac{C-(\mu+\delta)}{\sigma/\sqrt{n}}\right) \tag{106}$$

- Note: α and β usually change in opposite directions as n increases.

Review exercise 27, p. 55.

¶ **Choose the critical value and sample size,** C and n, so that:

- $\alpha =$ specified value, e.g. 0.02.
- $\beta \le$ specified value, e.g. 0.1.

Review exercise 28, p. 55.

¶ **Example: normal distribution.**

- Given: μ, δ and σ.
- Choose type-I error probability, α, say $\alpha = 0.02$.
- For any sample size n, the critical value, C, is found from eq.(103), p.28, as:

$$\Phi\left(\frac{C-\mu}{\sigma/\sqrt{n}}\right) = 1-\alpha \tag{107}$$

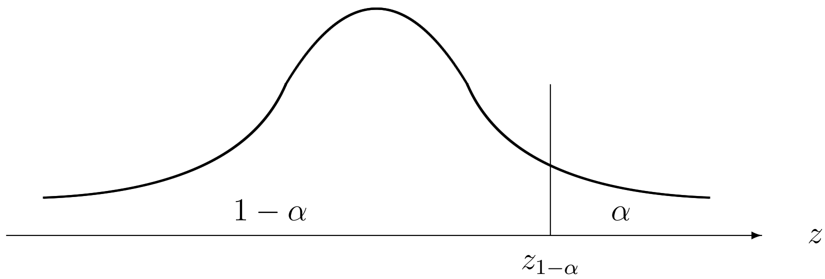


Figure 14: Sketch of probability density illustrating the critical value, eq.(110).

So:

$$\frac{C-\mu}{\sigma/\sqrt{n}} = z_{1-\alpha} = (1-\alpha)\text{th quantile of } \Phi \tag{108}$$

where the $(1-\alpha)$th quantile is defined (see fig. 14):

$$\text{Prob}(z \le z_{1-\alpha}) = 1-\alpha, \quad z \sim \mathcal{N}(0,1) \tag{109}$$

So:

$$C = \mu + \frac{z_{1-\alpha}\sigma}{\sqrt{n}} \tag{110}$$

• Now the type-II error probability, β, is, from eq.(106), p.28:

$$\beta = \Phi\left(\frac{C-(\mu+\delta)}{\sigma/\sqrt{n}}\right) \tag{111}$$

$$= \Phi\left(z_{1-\alpha} - \frac{\delta}{\sigma/\sqrt{n}}\right) \tag{112}$$

• Numerical example: $\mu = 0$, $\delta = 0.01$, $\sigma = 0.007$.
 ∘ Require $\alpha = 0.02$ so $z_{1-\alpha} = 2.05$.
 ∘ Table 6 shows n, C and β.

n	C, eq.(110)	$\frac{C-(\mu+\delta)}{\sigma/\sqrt{n}}$	β, eq.(111)
2	0.010147	0.029695	0.512
5	0.0064175	−1.14438	$1-0.8729=0.127$
10	0.0045379	-2.46754	$1-0.9934=0.0066$

Table 6: Sample size n, critical value C, and type-II error probability β.

 ∘ If we require $\beta \le 0.1$ then:
 — $n = 5$ is too small.
 — $n = 10$ is more than big enough.

• Suppose that $\delta > 0$. Then:

$$\Phi\left(z_{1-\alpha} - \frac{\delta}{\sigma/\sqrt{n}}\right) < \Phi(z_{1-\alpha}) \tag{113}$$

Thus, from eqs.(107) and (112):

$$\beta = \Phi\left(z_{1-\alpha} - \frac{\delta}{\sigma/\sqrt{n}}\right) < \Phi(z_{1-\alpha}) = 1-\alpha \tag{114}$$

That is:

$$\beta < 1-\alpha \tag{115}$$

Trade-off: small α means that β may be large.

8.2 Uncertain Effect Size and Variance

¶ **Effect size:**
• Effect size: $\Delta = \mu_0 - \mu_1$. Note: Δ (here) $= -\delta$ (section 8.1).
• Consider the upper-tail hypothesis: $\Delta < 0$.
• A similar derivation can be formulated for other cases.

¶ **The problem:**
• Assume normal distribution.
• We have an estimate of the effect size, $\widetilde{\Delta}$, but we are unsure how negative it really should be.
• We have an estimate $\widetilde{\sigma}$ of the population standard deviation but we are unconfident that this estimate is correct.

¶ **Fractional-error info-gap model:**

$$\mathcal{U}(h, \widetilde{\Delta}, \widetilde{\sigma}) = \Big\{ (\Delta, \sigma) : \quad (1+h)\widetilde{\Delta} \le \Delta \le \min[0,\ (1-h)\widetilde{\Delta}] \quad \max[0,\ (1-h)\widetilde{\sigma}] \le \sigma \le (1+h)\widetilde{\sigma} \Big\}, \quad h \ge 0 \tag{116}$$

¶ **Power =** $1 - \beta = 1-$ probability of type-II error, eq.(112), p.29:

$$\text{Power}(\Delta, \sigma, n) = 1 - \Phi\left(\frac{\Delta\sqrt{n}}{\sigma} + z_{1-\alpha}\right) \tag{117}$$

where $z_{1-\alpha}$ is the $(1-\alpha)$th quantile of the standard normal distribution.

¶ **Robustness** of sample size n, with requirement that the power be no less than $1 - \beta_{\text{c}}$:

$$\widehat{h}(n, \beta_{\text{c}}) = \max\left\{h : \left(\min_{(\Delta,\sigma)\in\mathcal{U}(h,\widetilde{\Delta},\widetilde{\sigma})} \text{Power}(\Delta, \sigma, n)\right) \ge 1 - \beta_{\text{c}}\right\} \tag{118}$$

¶ **Inner minimum** in eq.(118):

$$\mu(h) = \min_{(\Delta,\sigma)\in\mathcal{U}(h,\widetilde{\Delta},\widetilde{\sigma})} \text{Power}(\Delta, \sigma, n) \tag{119}$$

• $\mu(h)$ decreases as h increases: nesting of uncertainty sets.
• Robustness: greatest h such that $\mu(h) \ge 1 - \beta_{\text{c}}$.
• $\mu(h)$ is monotonic in h, so robustness is the max h satisfying $\mu(h) = 1 - \beta_{\text{c}}$.
• Plot of $\mu(h)$ vs. h is plot of $1 - \beta_{\text{c}}$ vs. $\widehat{\alpha}(n, \beta_{\text{c}})$.
• See fig. 15.

¶ **Derive the robustness function.**

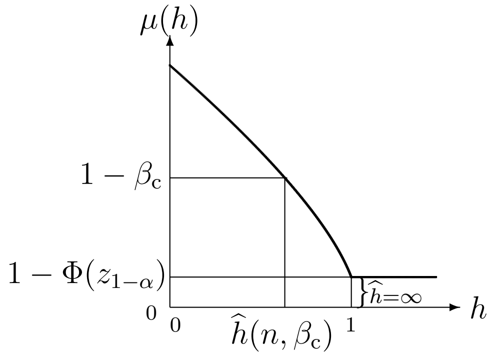


Figure 15: Illustration of the calculation of robustness.

• $\mu(h)$ occurs for the greatest allowed value of Δ/σ, which is negative and occurs when $\Delta = \min[0,\ (1-h)\widetilde{\Delta}]$ and when $\sigma = (1+h)\widetilde{\sigma}$.

• If $h \geq 1$ then:

◦ Power is minimized when $\Delta = 0$ so:

$$\text{Power}(\Delta, \sigma, n) = 1 - \Phi(z_{1-\alpha}) = \mu(h) \tag{120}$$

as in horizontal section of the curve in fig. 15.

◦ Robustness is infinite when $1-\beta_c < 1-\Phi(z_{1-\alpha})$.

◦ Thus very low demanded power (large β_c) implies very high robustness:

$$\widehat{h}(n, \beta_c) = \infty \quad \text{if } \beta_c > \Phi(z_{1-\alpha}) \tag{121}$$

¶ If $h < 1$: the robustness is the greatest value of h satisfying:

$$\Phi\left[\frac{(1-h)\widetilde{\Delta}\sqrt{n}}{(1+h)\widetilde{\sigma}} + z_{1-\alpha}\right] \leq \beta_c \tag{122}$$

• Let us denote by $q(\beta_c)$ the β_c quantile of the standard normal distribution:

$$\beta_c = \int_{-\infty}^{q(\beta_c)} \phi(x)\,\mathrm{d}x \tag{123}$$

• $q(\beta_c)$ increases from $-\infty$ to $+\infty$ as β_c increases from 0 to 1.

• The robustness is the greatest value of h satisfying:

$$\frac{(1-h)\widetilde{\Delta}\sqrt{n}}{(1+h)\widetilde{\sigma}} + z_{1-\alpha} \leq q(\beta_c) \tag{124}$$

• If:

$$\frac{\widetilde{\Delta}\sqrt{n}}{\widetilde{\sigma}} + z_{1-\alpha} > q(\beta_c) \tag{125}$$

then the robustness is zero for this value of β_c, and positive robustness is obtained only for greater values of β_c (lower power).

• Define:

$$\nu = \frac{q(\beta_{\rm c}) - z_{1-\alpha}}{\widetilde{\Delta}\sqrt{n}/\widetilde{\sigma}} \tag{126}$$

The robustness is positive only if $\nu < 1$ (recalling that $\widetilde{\Delta} < 0$).

• Now, solving eq.(124) (as an equality) for h, in the case that eq.(125) does not hold (that is, $\nu < 1$), we obtain the robustness:

$$\widehat{h}(n, \beta_{\rm c}) = \left\{ \begin{array}{ll} \dfrac{1-\nu}{1+\nu} & \text{if } \nu < 1 \\ 0 & \text{else} \end{array} \right. , \quad \text{if } \beta_{\rm c} \le \Phi(z_{1-\alpha}) \tag{127}$$

The complete robustness function is eqs.(121) and (127).

¶ Trade-off: Robustness vs. power.

Applying the chain rule for differentiation to eq.(127), one finds:

$$\frac{\partial \widehat{h}(n, \beta_{\rm c})}{\partial (1-\beta_{\rm c})} < 0 \tag{128}$$

The robustness $\widehat{h}(n, \beta_{\rm c})$ decreases as the demanded power, $1 - \beta_{\rm c}$, increases: high aspirations are vulnerable to uncertainty.

¶ Trade-off: Robustness vs. sample size.

One finds:

$$\frac{\partial \widehat{h}(n, \beta_{\rm c})}{\partial n} > 0 \tag{129}$$

Thus the robustness increases as the sample size increases.

8.3 Uncertain Sample PDF

8.3.1 Background

¶ Binary statistic test.

- $x =$ decision statistic.
- Distribution of x under H_1 equals distribution under H_0 shifted up by δ:

$$H_0 : \quad x \sim f(x) \tag{130}$$

$$H_1 : \quad x \sim f(x-\delta) \tag{131}$$

- We accept the null hypothesis iff $x \le C$.
- Determine:
  - Sample size, n.
  - Critical value, C.

¶ Definitions.

- $\alpha =$ level of significance = probability of type-I error (falsely reject H_0).
- $\beta(f) =$ probability of type-II error (falsely reject H_1) = $1-$ power.
- $\delta =$ non-negative effect size.
- $f(x) =$ pdf of decision statistic.
- $C =$ critical value.

$$1-\alpha = \int_{-\infty}^{C} f(x)\,\mathrm{d}x \tag{132}$$

$$\beta(f) = \int_{-\infty}^{C} f(x-\delta)\,\mathrm{d}x = \int_{-\infty}^{C-\delta} f(x)\,\mathrm{d}x = 1-\alpha-\int_{C-\delta}^{C} f(x)\,\mathrm{d}x \tag{133}$$

¶ Standard statistical approach.

- Known sampling distribution: $\widetilde{f}(x)$.
- $\widetilde{f}(x)$ depends on sample size (number of measurements.)
- Specify α and δ.
- Determine C and β from eqs.(132) and (133).
- Increase sample size until the power is adequate.

8.3.2 Info-gap Approach to Determining the Sample Size

¶ Approach.
- Sampling distribution is uncertain.
- Evaluate info-gap robustness of the estimated power.
- Determine sample size, n, so that adequate power is adequately robust.

¶ Info-gap model for pdf uncertainty: fractional-error.

$$\mathcal{U}(h,\widetilde{f}) = \left\{ f(x):\ f \in \mathcal{P},\ |f(x)-\widetilde{f}(x)| \le h\widetilde{f}(x) \right\}, \quad h \ge 0 \tag{134}$$

$\mathcal{P}$ is the set of all non-negative and normalized pdfs on the domain of x.

¶ Performance requirement.
- Power = $1-\beta$. Require large power; small β.
- $1-\beta_{\rm d} =$ demanded power.
- Analyst requires $\beta \le \beta_{\rm d}$.

¶ Robustness:

$$\widehat{h}(N,\beta_{\rm d}) = \max\left\{ h:\ \left(\max_{f\in\mathcal{U}(h,\widetilde{f})} \beta(f) \right) \le \beta_{\rm d} \right\} \tag{135}$$

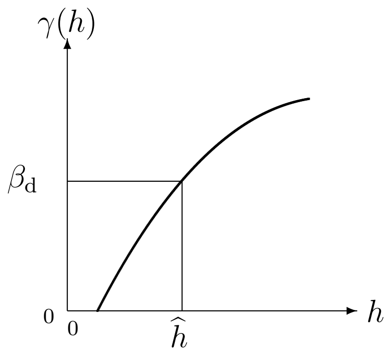


Figure 16: Illustration of the calculation of robustness.

¶ Evaluating robustness.
- Denote inner maximum in eq.(135) by $\gamma(h)$.
- Robustness is max h such that $\gamma(h) \le \beta_{\rm d}$.
- Uncertainty sets $\mathcal{U}(h,\widetilde{f})$ are nested with respect to h.
- Thus $\gamma(h)$ increases as h increases.
- Thus robustness is max h at which $\gamma(h) = \beta_{\rm d}$.
- $\gamma(h)$ is inverse of $\widehat{h}(N,\beta_{\rm d})$:

$$\gamma(h) = \beta_{\rm d} \quad \text{if and only if} \quad \widehat{h}(N,\beta_{\rm d}) = h \tag{136}$$

8.3.3 An Approximate Robustness for Small Effect Size

¶ **Special case,** very small effect size:

$$\delta \ll 1 \tag{137}$$

Derive approximate expression for robustness.

¶ Now eq.(133), p.33, can be approximated as:

$$\beta(f) = 1 - \alpha - f(C)\delta \tag{138}$$

¶ **Nominal critical value,** $\widetilde{C}$:
• $\widetilde{C}$ = $(1-\alpha)$th quantile of the best-estimated pdf $\widetilde{f}(x)$.
• $\widetilde{C}$ depends on sample size n.

¶ **Maximizing pdf.** The pdf in $\mathcal{U}(h, \widetilde{f})$ which maximizes β is very nearly:

$$\widehat{f}(x) = \begin{cases} \widetilde{f}(x) & \text{if } x < \widetilde{C} - \delta \\ (1-h)\widetilde{f}(x) & \text{if } x \in [\widetilde{C} - \delta,\ \widetilde{C}] \\ (1+wh)\widetilde{f}(x) & \text{if } x > \widetilde{C} \end{cases} \tag{139}$$

where w is a very small positive number which normalizes $\widehat{f}(x)$. That is, w is determined so that the decrement in $\widehat{f}$ in $[\widetilde{C} - \delta,\ \widetilde{C}]$ is compensated by the increment in $(\widetilde{C}, \infty)$:

$$wh[1 - \widetilde{F}(\widetilde{C})] = h\widetilde{f}(\widetilde{C})\delta \tag{140}$$

where $\widetilde{F}$ is the cumulative distribution function of $\widetilde{f}$.

¶ **Evaluating** $\gamma(h)$**.** $\beta(\widehat{f})$ in eq.(138) becomes:

$$\gamma(h) = \beta(\widehat{f}) = 1 - \alpha - (1-h)\widetilde{f}(\widetilde{C})\delta \tag{141}$$

Note: $\gamma(h)$ increases as h increases.

¶ **Evaluating robustness.**

Equate $\gamma(h)$ in eq.(141) to $\beta_{\rm d}$ and solve for h:

$$\widehat{h}(n, \beta_{\rm d}) = \begin{cases} 0 & \text{if } \beta_{\rm d} < 1 - \alpha - \widetilde{f}(\widetilde{C})\delta \\ \dfrac{\beta_{\rm d} - 1 + \alpha + \widetilde{f}(\widetilde{C})\delta}{\widetilde{f}(\widetilde{C})\delta} & \text{else} \end{cases} \tag{142}$$

• Robustness increases as $\beta_{\rm d}$ increases. Trade-off:
  high power $\Longleftrightarrow$ low robustness.
• Robustness is zero when $\beta_{\rm d}$ equals the nominal value, $\beta(\widetilde{f})$.
• Robustness depends on sample size through nominal critical value $\widetilde{C}$.
• This derivation is contingent on the assumption in eq.(137), p.35.

9 Tests of the Mean with Distributional Uncertainty

§ **Source:** Yakov Ben-Haim, 2008, Tests of the Mean with Distributional Uncertainty: An Info-Gap Approach, working paper.[2]

9.1 Distributional Uncertainty

§ **Statistical tests of the mean** depend on various assumptions about the data and population, such as:
• Normality.
• Random sampling: independent measurements with same instrument from same population which is uneffected by the measurement process.
• Stationarity of the sampled population.

§ **Distributional uncertainty:**
• Violations of assumptions about data and population, *unknown to the analyst*.
• E.g.:
◦ Non-normality.
◦ Sampling protocol varies. E.g., some observers are experts, some are not.
◦ Population evolves during the sample.
◦ Population is influenced by the sample.
• Examples:

◦ Franklin[3] uses a range of observational data from many different sources over the past 150 years—of varying and uncertain accuracy and reliability—to evaluate change in bird assemblages in northern Australia.

◦ McCarthy[4] uses museum collections to evaluate trends in marsupials and monotremes, recognizing that variable and uncertain collection efforts introduce uncertainties.

◦ Burgman *et al*[5] recognize that "collection frequencies will reflect changing trends in museum and herbarium collections", which introduces uncertainties in evaluating extinction threats based on historical development of collections.

◦ Stewart-Oaten *et al*[6] study tests of changes of a mean population property, before and after an impact, where the impact cannot be replicated (e.g., construction of a power plant). They note that data from such measurements "do not necessarily satisfy" the assumptions of standard tests. They state that "there is no panacea" for violation of test assumptions, and if the assumptions "are seriously wrong, alternative analyses are needed. This will often require a long time series of data." These authors discuss many sources of violation of test assumptions, stressing the importance of unknown skewness of distributions or correlations among measurements.

---

[2]Files: \papers\T-Test\ct03.tex and ttest07.tex.

[3]Franklin, Donald C., 1999, Evidence of disarray amongst granivorous bird assemblages in the savannas of northern Australia, a region of sparse human settlement, *Biological Conservation,* 90: 53–68.

[4]McCarthy, Michael A., 1998, Identifying declining and threatened species with museum data, *Biological Conservation,* 83: 9–17.

[5]Burgman, Mark A., Roger C. Grimson and Scott Ferson, 1995, Inferring threat from scientific collections, *Conservation Biology,* 9: 923–928.

[6]Stewart-Oaten, Allan, James R. Bence, and Craig W. Osenberg, 1992, Assessing effects of unreplicated perturbations: No simple solutions, *Ecology,* vol. 73, #4, pp.1396–1404.

§ **The problem:**

When violations are unknown and uncharacterized, the analyst cannot correct for them.

§ **Statistical tools exist for managing distributional uncertainty.**

- Careful test design.
- Non-parametric methods weaken some assumptions, e.g. normality.
  - These tests do assume random sampling, and usually are asymptotic.
  - They can be very sensitive to outliers.
- Given adequate data, one can model the data as a mixture of populations.
- Outliers can be managed using Jacknife or trimmed-means techniques.
- Method of M-estimates.

9.2 Info-Gap Representations of Distributional Uncertainty

§ θ is the test statistic. It may be a t statistic, but not necessarily.

§ **Tests of the mean:**

$$H_0: \qquad x \sim g(x) \tag{143}$$
$$H_1: \qquad x \sim g(x-\delta) \tag{144}$$

§ **Estimated pdfs.**
- Let $\widetilde{f}_i(\theta)$ denote the best guess of the pdf of the test statistic t, under hypothesis H_i.
- For instance:
  - ∘ If θ is the t statistic then $\widetilde{f}_0(\theta)$ is the t distribution with $n-1$ degrees of freedom.
  - ∘ $\widetilde{f}_1(\theta) = \widetilde{f}_0(\theta-\delta)$ where $\delta = (T_1 - T_0)/(s/\sqrt{n})$ is the shift between the two hypotheses.
  - ∘ Thus $\widetilde{f}_1(\theta)$ is formed by shifting $\widetilde{f}_0(\theta)$ to the right by δ.

§ **A fractional-error info-gap model:**

$$\mathcal{U}_i(h, \widetilde{f}_i) = \left\{ f(\theta): \ f(\theta) \in \mathcal{P}, \ |f(\theta) - \widetilde{f}_i(\theta)| \le h f_{\rm t}^\star, \ \forall \ \theta \right\}, \quad h \ge 0 \tag{145}$$

- $\mathcal{P}$ is the set of all normalized non-negative pdf's.
- $f_{\rm t}^\star$ is a normalization constant with units of probability density. For instance the mode:

$$f_{\rm t}^\star = \max_\theta \widetilde{f}(\theta) \tag{146}$$

If $\widetilde{f}(\theta)$ is a t distribution then $f_{\rm t}^\star = \widetilde{f}(0)$.

§ **A more restrictive fractional-error info-gap model:**

$$\mathcal{U}_i(h, \widetilde{f}_i) = \left\{ f(\theta): \ f(\theta) \in \mathcal{P}, \ |f(\theta) - \widetilde{f}_i(\theta)| \le h \widetilde{f}_i(\theta), \ \forall \ \theta \right\}, \quad h \ge 0 \tag{147}$$

- The variation on the tails dies out if $\widetilde{f}_i(\theta)$ becomes small on the tails, unlike eq.(145).

§ **Estimated cdfs.**
- Let $\widetilde{F}_i(\theta)$ denote the best guess of the cdf of the test statistic t, under hypothesis H_i.
- For instance:
  - ∘ If θ is the t statistic then $\widetilde{F}_0(\theta)$ is the t distribution with $n-1$ degrees of freedom for the statistic in eq.(168).
  - ∘ $\widetilde{F}_1(\theta) = \widetilde{F}_0(\theta - \delta)$ where $\delta = (T_1 - T_0)/(s/\sqrt{n})$.
  - ∘ Thus $\widetilde{F}_1(\theta)$ is formed by shifting $\widetilde{F}_0(\theta)$ to the right by δ.

§ **Uniform-bound info-gap model:**

$$\mathcal{U}_i(h) = \left\{ F(\theta): \ F(\theta) \in \mathcal{P}, \ |F(\theta) - \widetilde{F}_i(\theta)| \le h, \ \forall \ \theta \right\}, \quad h \ge 0 \tag{148}$$

where $\mathcal{P}$ is the set of all normalized non-negative cdf's.

9.3 Robustness Functions with CDF Uncertainty

§ This section is based on file \papers\T-Test\ct03.tex.

9.3.1 Binary Test: Formulation

§ **Data.**
- $X = \{x_1, \ldots, x_n\}$
- Not necessarily a random sample of any known distribution.

§ **Decision.** Two simple hypotheses about the population mean:

$$H_0: \quad \mu = T_0 \tag{149}$$
$$H_1: \quad \mu = T_1 \tag{150}$$

where each T_i is a specified number, and $T_1 > T_0$.

§ **Size and power.**
- θ is a statistic, for instance the t statistic.
- $F_i(\theta)$ is the cdf of θ under H_i.
- For any distribution $F(\theta)$, $q_\alpha(F)$ is the $(1-\alpha)$th quantile of $F(\theta)$:

$$1 - \alpha = F[q_\alpha(F)] \tag{151}$$

- We reject H_0 with significance α if:

$$\theta \geq q_\alpha(F_0) \tag{152}$$

- The size α, and power, $1-\beta$, are defined in:

$$1 - \alpha = F_0[q_\alpha(F_0)] \tag{153}$$
$$\beta = F_1[q_\alpha(F_0)] \tag{154}$$

- The size, α, is the probability of *falsely rejecting* the null hypothesis, H_0.
- $1-\alpha$ is the probability of correctly accepting H_0.
- The power, $1-\beta$, is the probability of *correctly rejecting* H_0.
- β is the probability of falsely rejecting H_1.

9.3.2 Robustnesses for Type I Errors

§ **Decision threshold.** Test of size $\alpha^\star$ which rejects H_0 when:

$$\theta \geq q_{\alpha^\star}(\widetilde{F}_0) \tag{155}$$

$\alpha^\star$: "nominal" size of the test, based on best-estimate of cdf under H_0, $\widetilde{F}_0$.

§ Note that:

$$\widetilde{F}_0[q_{\alpha^\star}(\widetilde{F}_0)] = 1 - \alpha^\star \tag{156}$$

§ **Robustness for falsely rejecting H_0:**
• Maximum horizon of uncertainty, h, at which the test at nominal size $\alpha^\star$ falsely rejects H_0 with probability no greater than α:

$$\widehat{h}_0(\alpha^\star, \alpha) = \max\left\{ h: \ \left(\min_{F \in \mathcal{U}_0(h)} F[q_{\alpha^\star}(\widetilde{F}_0)] \right) \geq 1 - \alpha \right\} \tag{157}$$

• We use the quantile $q_{\alpha^\star}(\widetilde{F}_0)$ because the test is implemented with the quantile of the best-guess distribution under H_0, $\widetilde{F}_0(\theta)$, and is of nominal size $\alpha^\star$.
• The actual size (probability of falsely rejecting H_0) is determined by the unknown true distribution under H_0, $F(\theta)$, which is info-gap-uncertain.

§ **Relation to type I error** (falsely rejecting H_0):
• $\widehat{h}_0(\alpha^\star, \alpha)$ is the greatest horizon of uncertainty up to which the probability of type I error is no greater than α.

§ **The Robustness,** $\widehat{h}_0(\alpha^\star, \alpha)$, for the info-gap model in eq.(148), is:

$$\widehat{h}_0(\alpha^\star, \alpha) = \alpha - \alpha^\star \tag{158}$$

or zero if this is negative.
• α is the **effective size**, while $\alpha^\star$ is the **nominal size**.
• For any choice of $\alpha^\star$, the robustness curve for type-I error, $\widehat{h}_0(\alpha^\star, \alpha)$ vs. α, is independent of the form of the test: t test, Wilcoxon signed-ranks test, etc.
• The implementation of the test, eq.(155), does depend on the type of test, through the value of the quantile $q_{\alpha^\star}(\widetilde{F}_0)$.

§ **Derivation of eq.(158).**
• Define the following step function:

$$V(x) = \begin{cases} 0, & \text{if } x < 0 \\ x, & \text{if } 0 \leq x \leq 1 \\ 1, & \text{else} \end{cases} \tag{159}$$

• Let $m_0(h)$ denote the inner minimum in eq.(157).
• The robustness, $\widehat{h}_0(\alpha^\star, \alpha)$, is the greatest non-negative h for which $m_0(h) = 1 - \alpha$.
• If there is no such h, then the robustness is zero.

- The inner min results when $F(\theta)$ is minimal at $q_{\alpha^\star}(\widetilde{F}_0)$, subject to membership in $\mathcal{U}_0(h)$.
- From the info-gap model in eq.(148) we find:

$$m_0(h) = V\left(\widetilde{F}_0[q_{\alpha^\star}(\widetilde{F}_0)] - h\right) = V(1 - \alpha^\star - h) \tag{160}$$

- Recall that $\widetilde{F}_0[q_{\alpha^\star}(\widetilde{F}_0)] = 1 - \alpha^\star$.
- The greatest value of h at which $m_0(h) = 1 - \alpha$ is the robustness, eq.(158).

9.3.3 Robustnesses for Type II Errors

§ **Robustness for falsely accepting H_0.**

• $\widehat{h}_1(\alpha^\star, \beta)$ is the greatest horizon of uncertainty up to which the probability of falsely accepting H_0, with a test of nominal size $\alpha^\star$, is no greater than β:

$$\widehat{h}_1(\alpha^\star, \beta) = \max \left\{ h : \ \left(\max_{F \in \mathcal{U}_1(h)} F[q_{\alpha^\star}(\widetilde{F}_0)] \right) \le \beta \right\} \tag{161}$$

§ $1 - \beta^\star$ is the nominal power:

$$1 - \beta^\star = 1 - \widetilde{F}_1[q_{\alpha^\star}(\widetilde{F}_0)] \tag{162}$$

• $\widehat{h}_1(\alpha^\star, \beta)$, for the info-gap model in eq.(148), is:

$$\widehat{h}_1(\alpha^\star, \beta) = 1 - \beta^\star - (1 - \beta) \tag{163}$$

or zero if this is negative.

• $1 - \beta$ as the **effective power**. $1 - \beta^\star$ is the **nominal power**.

• For any choice of $\alpha^\star$, $\widehat{h}_1(\alpha^\star, \beta)$ vs. β, depends on the form of the test, unlike for the type-I robustness. This is because the value of $\beta^\star$ depends on $\alpha^\star$ through the cdf's of the test statistic, $\widetilde{F}_0$ and $\widetilde{F}_1$.

§ **Derivation of eq.(163).**

• $m_1(h)$ denotes the inner maximum in eq.(161).

• The robustness, $\widehat{h}_1(\alpha^\star, \beta)$, is the greatest h at which $m_1(h) = \beta$.

• From the info-gap model in eq.(148), and using $V(x)$ in eq.(159):

$$m_1(h) = V\left(\widetilde{F}_1[q_{\alpha^\star}(\widetilde{F}_0)] + h\right) \tag{164}$$

Equating this to β and solving for h we find the robustness in eq.(163) with the aid of the expression for the nominal power in eq.(162).

9.3.4 Decisions and Judgments

§ **Two decisions, two judgments:**

• *Decide* on the nominal test size $\alpha^\star$ and the sample size n.

∘ Together these decisions determine the decision threshold $q_{\alpha^\star}(\widetilde{F}_0)$ in eq.(155), p.40.

• *Judge* what are reliable and acceptable values of effective size α and effective power $1-\beta$.

∘ Do this by considering $\widehat{h}_0(\alpha^\star, \alpha)$ and $\widehat{h}_1(\alpha^\star, \beta)$.

∘ α (size or level of significance) is the probability of falsely rejecting H_0.

∘ $1-\beta$ (power) is the probability of correctly rejecting H_0.

§ **Example: t test.**

• Test statistic, $\theta = (\overline{x} - T_0)(s/\sqrt{n})$. $\overline{x}$ is sample mean, s^2 is sample variance, and n is sample size.

• Estimated distribution under H_0, $\widetilde{F}_0(\theta)$, is the cdf of the t statistic with $n-1$ degrees of freedom.

• Estimated distribution under H_1 is $\widetilde{F}_1(\theta) = \widetilde{F}_0(\theta - \delta)$ where $\delta = (T_1 - T_0)/(s/\sqrt{n})$.

• True distributions under H_0 and H_1 are unknown; uncertainty is represented by info-gap model in eq.(148), p.38.

§ **No distributional uncertainty: no need for judgments:**

• $\alpha^\star$ is the actual size.

• Actual power, $1-\beta^\star$, is entirely determined by $\alpha^\star$ and n.

• Values of $\alpha^\star$ and $1-\beta^\star$ are shown in table 7.

• Power increases with increasing n at fixed $\alpha^\star$.

• Power increases with increasing $\alpha^\star$ at fixed n.

$\alpha^\star = 0.01$		$\alpha^\star = 0.05$	
n	$1-\beta^\star$	n	$1-\beta^\star$
5	0.1027	3	0.1784
7	0.3185	4	0.3736
9	0.5400	5	0.5390
12	0.7644	7	0.7457
31	0.9980	31	0.9997

Table 7: Size and power in the absence of distributional uncertainty.

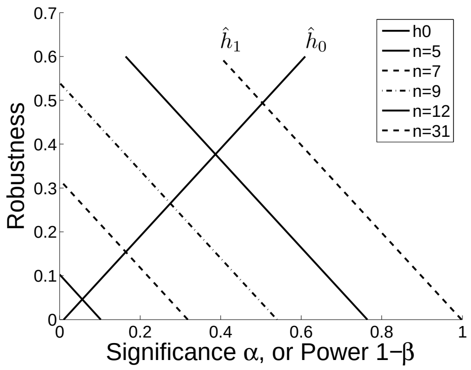


Figure 17: Robustness curves for the t test, $\widehat{h}_0(\alpha^\star, \alpha)$ for falsely rejecting H_0, and $\widehat{h}_1(\alpha^\star, \alpha)$ for falsely rejecting H_1. Nominal size is $\alpha^\star = 0.01$. $\widehat{h}_1(\alpha^\star, \alpha)$ calculated at 5 different sample sizes: $n = 5$, 7, 9, 12 and 31. $\delta = 1$.

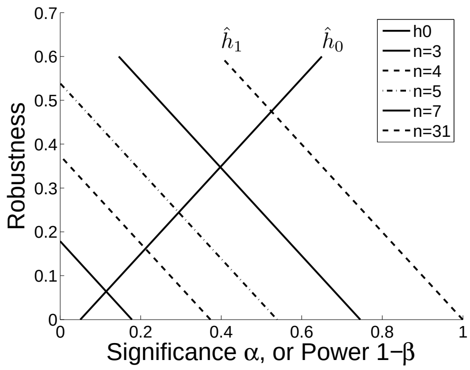


Figure 18: Robustness curves for the t test, $\widehat{h}_0(\alpha^\star, \alpha)$ for falsely rejecting H_0, and $\widehat{h}_1(\alpha^\star, \alpha)$ for falsely rejecting H_1. Nominal size is $\alpha^\star = 0.05$. $\widehat{h}_1(\alpha^\star, \alpha)$ calculated at 5 different sample sizes: $n = 3$, 4, 5, 7 and 31. $\delta = 1$.

§ **Robustness curves.** Figs. 17 and 18:

- $\widehat{h}_0(\alpha^\star, \alpha)$ vs. α (positive slope).
  - No robustness for nominal size: $\widehat{h}_0(\alpha^\star, \alpha^\star) = 0$.
  - Positive slope: Trade-off: robustness is exchanged for significance.
- $\widehat{h}_1(\alpha^\star, \beta)$ vs. $1 - \beta$ (negative slope).
  - No robustness for nominal power: $\widehat{h}_1(\alpha^\star, \beta^\star) = 0$.
  - Negative slope: Trade-off: robustness is exchanged for power.

§ **Judging reliable effective size, α:**

- The test designed for $\alpha^\star = 0.01$ will falsely reject H_0 with probability $\leq$ 0.05 if $F(\theta)$ differs from $\widetilde{F}_0(\theta)$ by no more than 0.04 in cumulative probability.
  - E.g., tails no heavier than 0.04 of total distribution.
  - E.g., outlying sub-population no larger than 0.04 of total distribution.

§ **Judging effective power, $1 - \beta$:**

- A test designed for size $\alpha^\star = 0.01$ with sample size $n = 9$ (dot-dash in fig. 17), has no robustness for power 0.54 (the horizontal intercept and nominal power).
- This test will falsely accept H_0 with probability of 0.44 if the actual cdf differs from the estimated cdf by no more than 0.1.

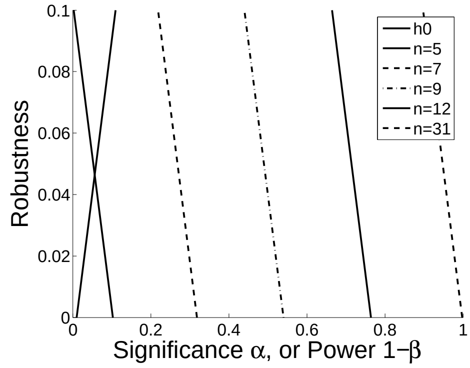


Figure 19: Same as fig. 17.

§ **Choosing the sample size, n.**

- Only type-II robustness is influenced by the sample size.
- The nominal and effective power:
  - ○ increase with increasing sample size,
  - ○ are influenced by the nominal size $\alpha^\star$.
- Choose n in light of the effective power and robustness which are needed.
- See fig. 19, which is expanded from fig. 17.

§ **Choosing the sample size, n, continued.**

- In fig. 19 consider nominal size $\alpha^\star = 0.01$.
- Judgment: effective size $\alpha = 0.05$ is adequate and reliable because $\widehat{h}_0(0.01, 0.05) = 0.04$.
- Apply this robustness to type II: Require $\widehat{h}_1(\alpha^\star, \beta) = 0.04$.
- From fig. 19: effective powers of 0.50, 0.72 and 0.96 for sample sizes 9, 12 and 31.
- Judgment: power of 0.50 is too small, so we require a sample larger than $n = 9$.
- Judgment: if power of 0.72 is adequate then we adopt a sample of size 12.
- Choosing a sample of size 31 would result in power of 0.96.

§ **Judgments of robustness: how much robustness is needed?**

- Robustness has units of probability (in this example).
- Thus judge adequate robustness probabilistically.
- This not necessary: analogical inference.

§ **Choosing the sample size, n, continued.**

- Previously we required $\widehat{h}_0(\alpha^\star, \alpha) = \widehat{h}_1(\alpha^\star, \beta)$.
- This is not necessary. We can make separate judgments for type I and type II robustnesses.

9.4 Robustness Functions with PDF Uncertainty: Definitions

§ This section and the next are based on file \papers\T-Test\ttest07.tex.

9.4.1 Binary Test: Formulation

§ **Data.**
• $X = \{x_1, \ldots, x_n\}$
• Not necessarily a random sample of any known distribution.

§ **Decision.** Two simple hypotheses about the population mean:

$$H_0 : \quad \mu = T_0 \tag{165}$$
$$H_1 : \quad \mu = T_1 \tag{166}$$

where each T_i is a specified number, and $T_1 > T_0$.

§ **Sample mean and variance:**

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{167}$$

§ **The t statistic for testing H_0 is:**

$$t = \frac{\overline{x} - T_0}{s/\sqrt{n}} \tag{168}$$

which has a t distribution with $n-1$ degrees of freedom under H_0 (in the absence of distributional uncertainty).

§ **Size and power of the test.**
• Let $f_i(t)$ denote the probability density of t under H_i.
• For any density $f(t)$, let $q_\alpha(f)$ denote the $(1-\alpha)$th quantile of $f(t)$:

$$\int_{-\infty}^{q_\alpha(f)} f(t)\,\mathrm{d}t = 1 - \alpha \tag{169}$$

• We reject H_0 with significance α if:

$$t \geq q_\alpha(f_0) \tag{170}$$

• The size α, and power, $1-\beta$, are defined in:

$$1 - \alpha = \int_{-\infty}^{q_\alpha(f_0)} f_0(t)\,\mathrm{d}t \tag{171}$$
$$\beta = \int_{-\infty}^{q_\alpha(f_0)} f_1(t)\,\mathrm{d}t \tag{172}$$

◦ α is the probability of falsely rejecting the null hypothesis, H_0.
◦ $1-\beta$ is the probability of correctly rejecting H_0.

9.4.2 Robustness for Falsely Rejecting H_0. (Type I Error.)

• Consider a t test of size $\alpha^\star$, which rejects H_0 when:

$$t > q_{\alpha^\star}(\widetilde{f}_0) \tag{173}$$

• The robustness is the maximum horizon of uncertainty, h, up to which the t test at size $\alpha^\star$ falsely rejects H_0 with probability no greater than α:

$$\widehat{h}_0(t, \alpha^\star, \alpha) = \max\left\{ h : \left(\min_{f \in \mathcal{U}_0(h, \widetilde{f}_0)} \int_{-\infty}^{q_{\alpha^\star}(\widetilde{f}_0)} f(t)\, \mathrm{d}t \right) \geq 1 - \alpha \right\} \tag{174}$$

• $q_{\alpha^\star}(\widetilde{f}_0)$: the test is implemented with the quantile of the best-guess distribution under H_0, $\widetilde{f}_0$, and is of nominal size $\alpha^\star$.
• Actual size (probability of falsely rejecting H_0) is determined by the unknown true distribution under H_0, f.
• The inverse of $\widehat{h}_0(t, \alpha^\star, \alpha)$ is defined as:

$$m_0^{\mathrm{t}}(h, \alpha^\star) = 1 - \alpha \quad \text{if and only if} \quad \widehat{h}_0(t, \alpha^\star, \alpha) = h \tag{175}$$

• An explicit expression for the inverse of $\widehat{h}_0(t, \alpha^\star, \alpha)$ is:[7]

$$m_0^{\mathrm{t}}(h, \alpha) = [c_1(h) - c_2(h)]hf_{\mathrm{t}}^\star + \widetilde{F}_0[c_2(h)] - \widetilde{F}_0[c_1(h)] \tag{176}$$

where:

$$c_1(h) = -\widetilde{f}_0^{-1}(hf_{\mathrm{t}}^\star) \tag{177}$$
$$c_2(h) = \min[\widetilde{f}_0^{-1}(hf_{\mathrm{t}}^\star),\ q_{\alpha^\star}(\widetilde{f}_0)] \tag{178}$$

∘ $\widetilde{f}_0(t)$ is the pdf of the t variate with $n-1$ dofs.
∘ $\widetilde{F}_0(t)$ is the cumulative distribution function of the t variate with $n-1$ dofs.
∘ $\widetilde{f}_0^{-1}(h)$ is the inverse of $\widetilde{f}_0(t)$ for $t \geq 0$.
Thus $-\widetilde{f}_0^{-1}(h)$ is the smallest value of t at which $\widetilde{f}_0(t) = h$.
So $-\widetilde{f}_0^{-1}(0) = -\infty$ and $\widetilde{f}_0^{-1}[\widetilde{f}_0(0)] = 0$.

§ **Type I error (falsely rejecting H_0).**
∘ $\widehat{h}_0(t, \alpha^\star, \alpha)$ is the greatest horizon of uncertainty at which:
the probability of type I error is no greater than α.
∘ The test is implemented so that the probability of type I error is no greater than $\alpha^\star$ assuming no distributional uncertainty.

[7]Yakov Ben-Haim, 2008, Tests of the Mean with Distributional Uncertainty: An Info-Gap Approach, working paper. Appendix A.

9.4.3 Robustness for Falsely Accepting H_0. (Type II Error.)

• Consider, as before, a t test of size $\alpha^\star$, which rejects H_0 when:

$$t > q_{\alpha^\star}(\widetilde{f}_0) \tag{179}$$

• The robustness is the greatest horizon of uncertainty up to which the probability of falsely accepting H_0, with a t test of size $\alpha^\star$, is no greater than β:

$$\widehat{h}_1(t, \alpha^\star, \beta) = \max\left\{h : \left(\max_{f\in\mathcal{U}_1(h,\widetilde{f}_1)} \int_{-\infty}^{q_{\alpha^\star}(\widetilde{f}_0)} f(t)\,\mathrm{d}t\right) \le \beta\right\} \tag{180}$$

• An explicit expression for the inverse of $\widehat{h}_1(t, \alpha^\star, \beta)$ is:[8]

$$M_1^{\mathrm{t}}(h, \alpha^\star) = 1 + (c_4 - c_3)hf_{\mathrm{t}}^\star - \widetilde{F}_1(c_4) + \widetilde{F}_1(c_3) \tag{181}$$

where:

$$c_3(h) = q_{\alpha^\star}(\widetilde{f}_0) \tag{182}$$

$$c_4(h) = \max[\widetilde{f}_1^{-1}(hf_{\mathrm{t}}^\star),\ q_{\alpha^\star}(\widetilde{f}_0)] \tag{183}$$

◦ We have assumed that $q_{\alpha^\star}(\widetilde{f}_0) \ge \delta$, which in practice will always hold.
◦ $\widetilde{f}_1(t) = \widetilde{f}_0(t-\delta)$.
◦ $\widetilde{F}_1(t)$ is the cumulative distribution function of the $\widetilde{f}_1(t)$.
◦ $\widetilde{f}_1^{-1}(h)$ is the inverse of $\widetilde{f}_1(t)$ for $t \ge \delta$.

• A plot of h vs. $M_1^{\mathrm{t}}(h, \alpha^\star)$ is identical to a plot of $\widehat{h}_1(t, \alpha^\star, \beta)$ vs. β.

§ **Nominal power.**

• Let $1 - \beta^\star$ be the nominal power:

$$\beta^\star = \int_{-\infty}^{q_{\alpha^\star}(\widetilde{f}_0)} \widetilde{f}_1(t)\,\mathrm{d}t \tag{184}$$

From the contraction and nesting axioms we recognize that $\widehat{h}_1(t, \alpha^\star, \beta^\star) = 0$ and $\widehat{h}_1(t, \alpha^\star, \beta) > 0$ only for $\beta > \beta^\star$.

[8] Yakov Ben-Haim, 2008, Tests of the Mean with Distributional Uncertainty: An Info-Gap Approach, working paper. Appendix B.

9.5 Robustness with PDF Uncertainty: Numerical Examples

9.5.1 Robustness for Type-I Error

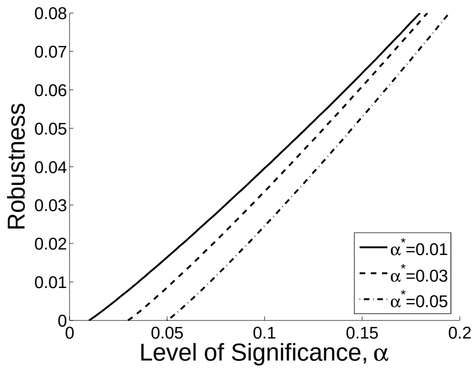


Figure 20: Robustness curves for the t test, $\widehat{h}_0(t, \alpha^\star, \alpha)$, for falsely rejecting H_0, at three design sizes, $\alpha^\star = 0.01$, 0.03 and 0.05. Sample size $n = 17$. $f_{\rm t}^\star = \max_t \widetilde{f}_0(t)$.

§ **Trade-off:**

- Robustness vs. level of significance.
- Zero robustness at nominal level of significance.

§ **What does $\widehat{h}_0(t, \alpha^\star, \alpha) = 0.02$ mean?**

- The true pdf, $f(t)$, can deviate from $\widetilde{f}_0(t)$ by a 'bump' (or dimple) no larger than 0.02$f_{\rm t}^\star$ if size α is not to be exceeded.
- This is a small bump: the tail of $\widetilde{f}_0(t)$ becomes as thin as 0.02$f_{\rm t}^\star$ at about 3 σ's from the mean.
- So, $\widehat{h}(t, \alpha^\star, \alpha) = 0.02$ might be sufficient robustness only if immunity to small deviations on the far tails is sufficient.

9.5.2 Robustness for Type-II Error

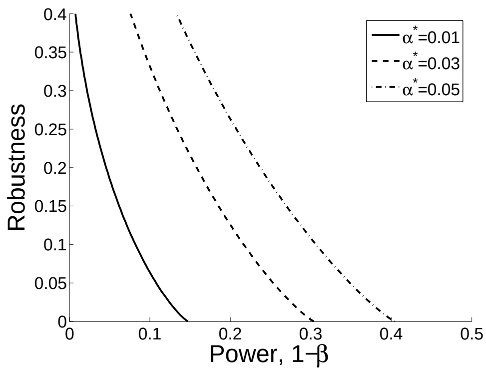


Figure 21: Robustness curves for the t test, $\widehat{h}_1(t, \alpha^\star, \beta)$, for correctly rejecting H_0, at three design sizes, $\alpha^\star = 0.01$, 0.03 and 0.05. Sample size $n = 17$. $f_{\rm t}^\star = \max_t \widetilde{f}_0(t)$.

§ **Trade-off:**

- Robustness vs. power of the test.

- Zero robustness at nominal power.

10 Accelerated Lifetime Testing: Simple Case

10.1 Formulation

§ **Lifetime testing:** Measure MTTF or other statistical characterization of a system under operating conditions.[9]

§ **Accelerated lifetime testing:** Measure MTTF or other statistical characterization of a system under conditions which are *more stressful* than ordinary operating conditions. Then *deduce lifetime* under ordinary conditions.

§ **Lifetime** of a device is denoted ℓ, which depends on the "stress" which the system is subject to: $\ell(s)$.

§ **Linear lifetime-stress model:**
- We adopt a piece-wise linear model:

$$\ell_{\rm m}(s,c) = \begin{cases} (s-s_0)c & \text{if } s \le s_0 \\ 0 & \text{if } s \ge s_0 \end{cases} \tag{185}$$

- $c < 0$.
- s_0 is known.
- That is, $\ell = 0$ for stress greater than s_0.
- Lifetime increases as stress decreases below s_0.

§ **Data:** we have measured (estimated) ℓ at stress $s_1 < s_0$: $\ell(s_1)$ is known.

§ **Best-estimated model:** Given the data, we estimate c:

$$\widehat{c} = \frac{\ell(s_1)}{s_1 - s_0} \tag{186}$$

§ **Requirement:** Estimate $\ell(s)$ for $s_2 < s_1$.

10.2 Uncertainty and Robustness

§ **Uncertain information:** We expect that:

$$\ell(s_2) > \ell_{\rm m}(s_2, \widehat{c}) \tag{187}$$

- Lifetime at low stress, s_2, should be greater than linear prediction.
- We don't know how much greater.

§ **Info-gap model of uncertainty:**

$$\mathcal{U}(h) = \left\{ \ell(s_2) : \ 0 \le \ell(s_2) - \ell_{\rm m}(s_2, \widehat{c}) \le h \right\}, \quad h \ge 0 \tag{188}$$

[9]This example is programmed in GapZapper: Domain: Statistics, Application: Accel-Lifetime-Test-Simple.

§ **Performance function:**
• Consider linear model with coefficient c. Squared error is:

$$E^2(c) = [\ell(s_1) - \ell_{\mathrm{m}}(s_1, c)]^2 + [\ell(s_2) - \ell_{\mathrm{m}}(s_2, c)]^2 \tag{189}$$

§ **Performance requirement:**

$$E^2(c) \le E_{\mathrm{c}}^2 \tag{190}$$

§ **Robustness of linear model $\ell_{\mathrm{m}}(s, c)$:**

$$\widehat{h}(c, E_{\mathrm{c}}) = \max\left\{ h : \left(\max_{\ell(s_2)\in\mathcal{U}(h)} E^2(c) \right) \le E_{\mathrm{c}}^2 \right\} \tag{191}$$

10.3 Evaluating the Robustness

§ We begin by evaluating the inverse of the robustness. We then invert this.

§ Let $\mu(h)$ denote the inner maximum in the robustness, eq.(191). This is the **inverse of the robustness:**

$$\mu(h) = E_{\mathrm{c}}^2 \quad \text{implies} \quad \widehat{h}(c, E_{\mathrm{c}}) = h \tag{192}$$

Equivalently:

$$\sqrt{\mu(h)} = E_{\mathrm{c}} \quad \text{implies} \quad \widehat{h}(c, E_{\mathrm{c}}) = h \tag{193}$$

• Plot of h vs $\sqrt{\mu(h)}$ is the same as $\widehat{h}(c, E_{\mathrm{c}})$ vs E_{c}.

§ **Extreme values of** $\ell(s_2)$ at horizon of uncertainty h. From info-gap model of eq.(188):

$$\ell_{\mathrm{m}}(s_2, \widehat{c}) \le \ell(s_2) \le \ell_{\mathrm{m}}(s_2, \widehat{c}) + h \tag{194}$$

§ **Evaluating** $\mu(h)$:
• $\mu(h)$ occurs at one of the extreme values of $\ell(s_2)$.
• That is, $\mu(h)$ is the greater of the following two values:

$$\mu_1 = [\ell(s_1) - \ell_{\mathrm{m}}(s_1, c)]^2 + [\ell_{\mathrm{m}}(s_2, \widehat{c}) - \ell_{\mathrm{m}}(s_2, c)]^2 \tag{195}$$

$$\mu_2(h) = [\ell(s_1) - \ell_{\mathrm{m}}(s_1, c)]^2 + [\ell_{\mathrm{m}}(s_2, \widehat{c}) + h - \ell_{\mathrm{m}}(s_2, c)]^2 \tag{196}$$

• **Assumption:** Given our uncertain information, eq.(187), we will only consider slopes c which are steeper than $\widehat{c}$:

$$c \le \widehat{c} < 0 \tag{197}$$

Recall that s_0 is known with certainty.
• Thus:

$$\ell_{\mathrm{m}}(s_2, c) \ge \ell_{\mathrm{m}}(s_2, \widehat{c}) \tag{198}$$

• Hence the inverse robustness is:

$$\mu(h) = \begin{cases} \mu_1 & \text{if } h < 2[\ell_{\mathrm{m}}(s_2, c) - \ell_{\mathrm{m}}(s_2, \widehat{c})] \\ \mu_2(h) & \text{else} \end{cases} \tag{199}$$

• When $c = \widehat{c}$ then $\mu(h) = \mu_2(h)$. Also note that from the definition of $\widehat{c}$ one finds that $\ell(s_1) = \ell_{\rm m}(s_1, \widehat{c})$. Hence:

$$\mu_2(h) = h^2 \tag{200}$$

So:

$$\widehat{h}(\widehat{c}, E_{\rm c}) = E_{\rm c} \tag{201}$$

Generally, we see that eq.(199) can be inverted and, together with eq.(195) and (196), we obtain an explicit expression for the robustness. Set $\mu_2(h) = E_{\rm c}^2$ in eq.(199) and solve for h to obtain:

$$\widehat{h}(c, E_{\rm c}) = \begin{cases} 0 & \text{if } E_{\rm c} \le \sqrt{\mu_1} \\ \sqrt{E_{\rm c}^2 - [\ell(s_1) - \ell_{\rm m}(s_1, c)]^2} - \ell_{\rm m}(s_2, \widehat{c}) + \ell_{\rm m}(s_2, c) & \text{else} \end{cases} \tag{202}$$

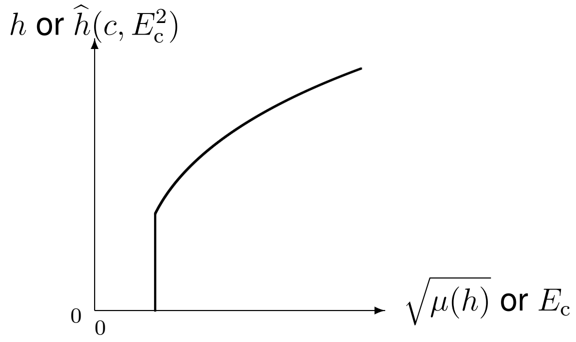


Figure 22: Illustration of the robustness based on eq.(199).

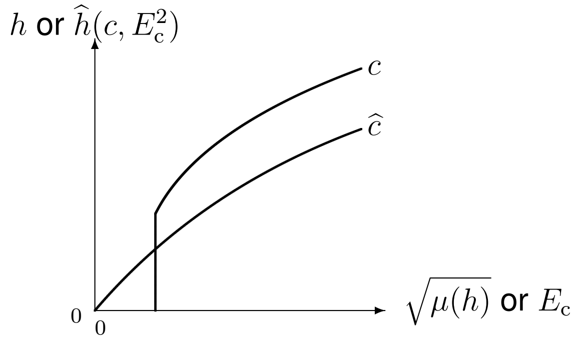


Figure 23: Illustration of the robustness based on eq.(199).

11 Review Exercises

§ The exercises in this section are not homework problems, and they do not entitle the student to credit. They will assist the student to master the material in the lecture and are highly recommended for review and self-study.

1. Explain that the variance, eq.(7), p.4, can we written as:

$$\mathrm{var}(\overline{x}) = \mathrm{E}(x^2) - \mathrm{E}(x)^2 \tag{203}$$

2. What is it significant and important that theorem 1, p.3, does not depend on the probability distribution of the observations?
3. Does eq.(10), p.4, depend on the probability distribution of the observations?
4. The central limit theorem, p.4, explains the widespread (but not universal) occurrence of normal distributions. (Is this why the normal distribution is called "normal"?) Think up some examples of natural, social, or other phenomena that are described by normal distributions.
5. Why is the hypothesis test in eqs.(12) and (13), p.5, called a "two-tailed" test? Why is the hypothesis test in eqs.(14) and (15), p.5, called a "one-tailed" test?
6. Why is the level of confidence, eq.(16), p.5, defined as "the probability of obtaining a result *at least as extreme as* the observed result", rather than as "the probability of obtaining a result *equal to* the observed result"?
7. Explain eq.(18), p.6: where does the "standard normal" distribution, $\mathcal{N}(0, 1)$, come from?
8. Why is $N \geq 25$ a "rough number" on p.7? Why can't we state a precise value for N above which the normal distribution applies exactly? For what type of distribution would you need $N \gg 25$ or $N \ll 25$?
9. The probability of 0.14, in eq.(30), p.8, may not sound small to everybody. How do you decide what is small, medium, huge, big enough, etc?
10. Do you agree with the assertion, on p.9, that $\alpha = 0.14$ is "not very convincing"? Why?
11. Will the sequential procedure in table 1, p.9, always result in α *decreasing* and thus leading to rejection of H_0? What would the table look like if really H_0 should be rejected?
12. When would you choose the alternative hypothesis in eq.(33) or (34) rather than (32), p.10?
13. Why can't we answer the question following eq.(36), p.10, under hypothesis H_1?
14. Explain eqs.(43) and (44), p.11.
15. Why is eq.(46), p.11, true?
16. $0.21 > 0.15$, so why isn't the answer to the question following eq.(51), p.13, obvious?
17. Regarding the conclusion that the engine is running hot, following eq.(55), p.15: Didn't we already know this from the fact that $0.21 > 0.15$? What additional insight does the χ^2 test provide (if any)?
18. The χ^2 test is used to test "categorical" data: events that occur in different types, classes, or categories, as distinct from events that result in a real number. The tests in sections 4, p.13, 5, p.16, and 6, p.18, are very different, though they are all χ^2 tests. What are the categories in these tests?

19. What is the difference between $p_{\rm u}$ and $p_{\rm a}$ in eq.(80), p.22? Why isn't "unacceptable" anything greater than "acceptable"? What deeper problem of meaning and judgment is this distinction trying to grapple with?
20. Are the two types of errors on p.22 the *only* errors that one can make? Are they equally severe?
21. What assumptions underlie the binomial distribution in eq.(82), p.22?
22. Why is $P_{\rm I}$ in eq.(86), p.23, call the consumer's risk? Why is $P_{\rm II}$ in eq.(87), p.23, call the producer's risk?
23. In the example on p.24, is a sample size $N = 200$ better or worse than a sample size of $N = 100$? Why?
24. An infinite sample size is ideal, as shown in fig. 11, p.25, but this is usually not feasible. What sample size is big enough? Why does it matter who decides?
25. In fig. 12, p.26, explain why $\delta > 0$ implies (1) pdf shifts to the right and (2) is represented by $g(t-\delta)$ rather than $g(t+\delta)$.
26. Explain eq.(93), p.27. (Recall theorems in section 1).
27. Why do α and β in eqs.(101)–(106), p.28, change in opposite directions as n increases? What is the significance of this?
28. Why the different treatment of α and β on p.28: $\alpha = 0.02$ and $\beta \le 0.1$?