Statistics

Week 3: Sampling Distributions (Chapter 5), Estimation (Chapter 6)

ESD, SUTD

Term 5, 2017

SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

Established in collaboration with MIT

Homework assignment 1

Due 1:00 pm, 14 Feb. Submit on *eDimension*. Show working.

Hints:

Q4(c): you don't need to use up all the portions; there are multiple sensible designs.

Q5(b): you can draw the box plot by hand or in *R*.

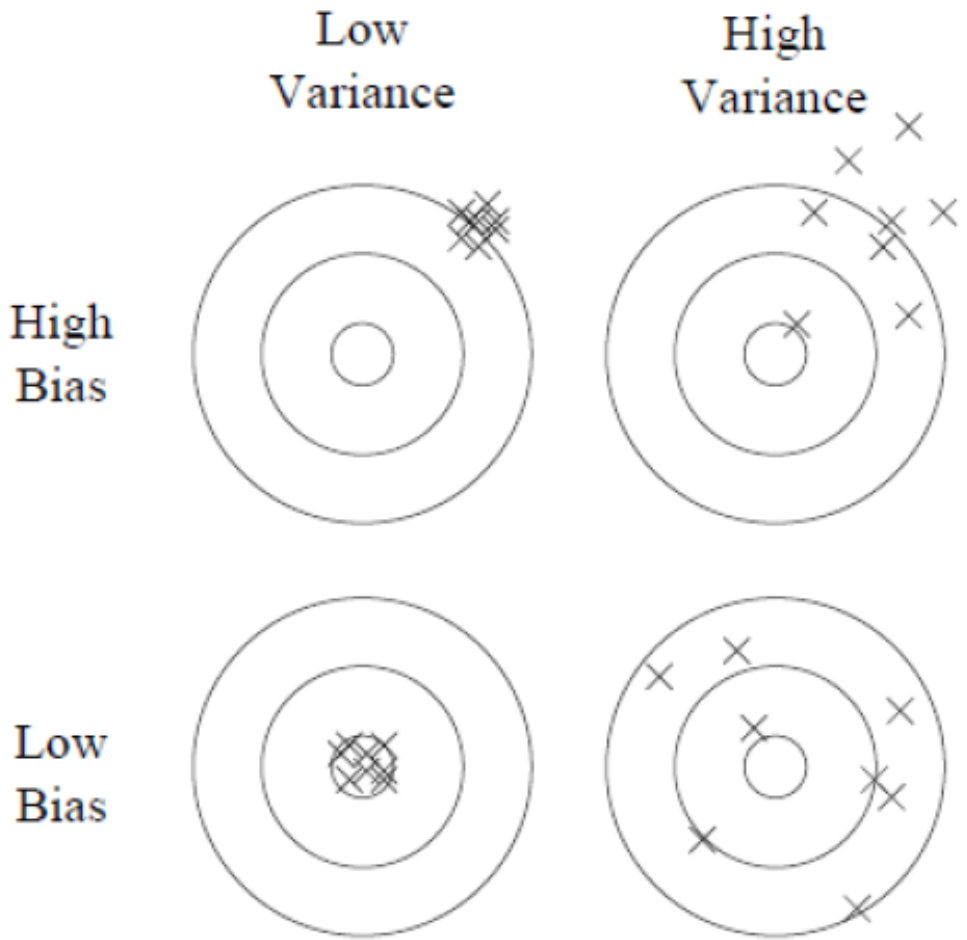
Q7: assume that birthdays are uniformly distributed, and that there are 52 weeks in a year.

Q8(a): we've done a similar calculation in class.

Q9: this question is harder; be creative.

Outline

Bias and variance of an estimator



The relationship between bias and variance (of an estimator) is analogous to the relationship between *accuracy* and *precision*.

Exercise (adapted from 2015 exam)

Let the iid random variables X_1, X_2, X_3 be drawn from a distribution with mean μ and variance σ^2.

(1) If the estimator for μ,

$$\hat{\mu} = c_1 X_1 + c_2 X_2 + c_3 X_3$$

is unbiased, then what relation must the constants c_1, c_2, c_3 satisfy?

(2) Find, with proof, the values of c_1, c_2, c_3 such that $\mathrm{Var}(\hat{\mu})$ is minimized.

Answers

(1) $\mathsf{E}(\hat{\mu}) = c_1\,\mathsf{E}(X_1) + c_2\,\mathsf{E}(X_2) + c_3\,\mathsf{E}(X_3) = (c_1 + c_2 + c_3)\mu$, so $c_1 + c_2 + c_3 = 1$.

(2) $\mathsf{Var}(\hat{\mu}) = (c_1^2 + c_2^2 + c_3^2)\sigma^2$.

To minimize $c_1^2 + c_2^2 + c_3^2$ subject to the constraint $c_1 + c_2 + c_3 = 1$, we may use Lagrange multipliers (among other methods).

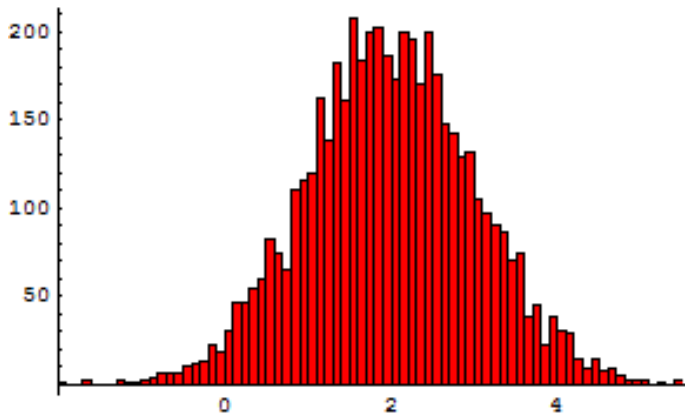
The minimum is achieved when $c_1 = c_2 = c_3 = \dfrac{1}{3}$, that is, when $\hat{\mu} = \bar{x}$.

Outline

Quality control – toy example

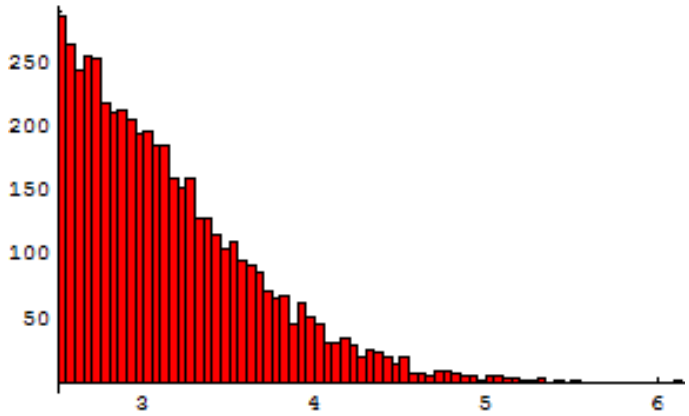
You ask a supplier to give you 5000 of their products, selected at random, to test if the mean weight is 2.5. You know that the weight is normally distributed with variance σ^2.

You plot the weights on a *histogram*.

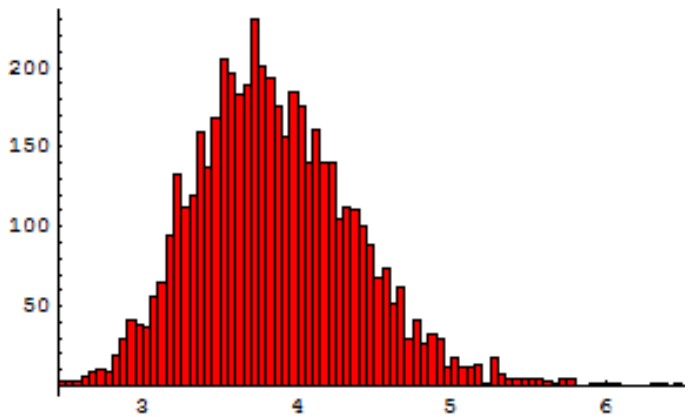


So they fail. The next month, they again give you 5000 products, but not completely randomly: only products with weight > 2.5 are selected.

Their scheming is discovered using another histogram.



The month after, they grow more cunning, and each product given actually has the *maximum* weight out of a batch of 10.



The histogram looks normal. . . How to uncover their cheating?

Answer: use a Q-Q plot, or study the *variance*.

Chi-squared distribution

The CLT gives the approximate distribution for the sample mean when the sample size is large. Unfortunately, there is no such theorem for the sample variance drawn from an arbitrary distribution.

However, if the distribution is *normal*, then the behaviour of the sample variance s^2 is well-understood, in terms of the **chi-squared distribution**.

Chi-squared random variable

A chi-squared random variable with n *degrees of freedom*, denoted by χ^2_n, is defined as the sum of squares of n iid standard normal random variables.

Chi-squared – pdf

The probability density function of a chi-squared random variable with n degrees of freedom is given by

$$\frac{1}{2^{n/2}\,\Gamma(n/2)}\,x^{n/2-1}e^{-x/2}.$$

- Note 1: $\Gamma(n)$ denotes the Gamma function, which is a continuous interpolation of the factorial function. $\Gamma(n+1) = n\,\Gamma(n)$, with $\Gamma(1) = 1$ and $\Gamma(1/2) = \sqrt{\pi}$.
- Note 2: although we will not directly work with the above formula, the existence of a closed form for the pdf means it is easy to implement in computer programs, and hence useful in actual calculations.

Chi-squared – proof

Step 1: let Z be a standard normal r. v. Let F be the cdf of Z^2 and Φ be the cdf of Z. Then

$$F(x) = P(Z^2 \leq x) = P(-\sqrt{x} \leq Z \leq \sqrt{x}) = \Phi(\sqrt{x}) - \Phi(-\sqrt{x}).$$

Differentiating the first term and the last term with respect to x, we find that the pdf of Z^2 is

$$f(x) = \frac{1}{\sqrt{2\pi x}} e^{-x/2}.$$

Step 2: from the above pdf, we find that the *moment generating function* of Z^2 is

$$M_{Z^2}(t) = (1-2t)^{-1/2}.$$

On the other hand, from the pdf for χ^2_n, the corresponding mgf is $(1-2t)^{-n/2}$.

Step 3: use the fact that when X and Y are independent, $M_{X+Y}(t) = M_X(t)M_Y(t)$.

Chi-squared and variance

Let $X_1, \ldots, X_n$ be iid normal random variables with mean μ and variance σ^2, then

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}.$$

The proof is similar to the calculation of $\mathrm{E}(s^2)$

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 = \sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 + \sum_{i=1}^{n} \left(\frac{\bar{X} - \mu}{\sigma}\right)^2 + 0$$

$$= \frac{(n-1)s^2}{\sigma^2} + \left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}}\right)^2.$$

LHS $\sim \chi^2_n$, last term $\sim \chi^2_1$.

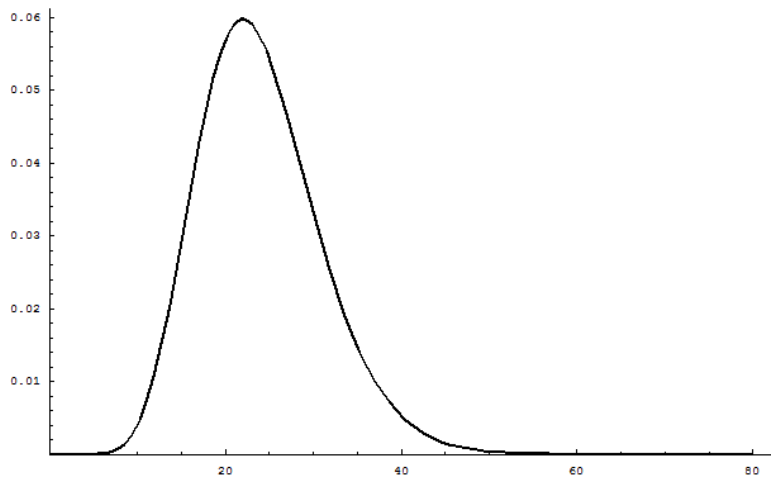
Also, it follows from the variance of χ^2_{n-1} that $\mathrm{Var}(s^2) = \dfrac{2\sigma^4}{n-1}$.

Chi-squared – exercise

The waiting times in a bank are normally distributed with a standard deviation of 8.2 minutes. What is the probability that for a random sample of 25 customers, the sample standard deviation is greater than 10 minutes?

Use the *Excel* command `chisq.dist`.

Answer: 0.0588



t-distribution

Let $X_1, \ldots, X_n$ be iid normal random variables with mean μ and variance σ^2. Let the random variable T_{n-1} be

$$T_{n-1} = \frac{\bar{X} - \mu}{s/\sqrt{n}} = \frac{Z}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}.$$

Definition: T_{n-1} follows a **Student t-distribution** with $(n-1)$ degrees of freedom.

The t-distribution is symmetric and bell-shaped, but has heavier tails than the standard normal distribution; it converges to the standard normal as $n \to \infty$.

t-distribution: formula and history

Using transformation of random variables (from Probability), we can show that the pdf for T_{n-1} is

$$f_{n-1}(t) = \frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\,\Gamma(\frac{n-1}{2})}\left(1+\frac{t^2}{n-1}\right)^{-\frac{n}{2}}.$$

The t-distribution was popularized by William Gosset, who was a researcher at Guinness Brewery and published under the pen name 'Student', since employees of the brewery were forbidden to publish papers lest they revealed trade secrets.