MIE 1807
Principles of Measurement

2017-02-07
Neil Montgomery

Lecture notes:
https://github.com/mie1807-winter-2017

Limit theorems not always useful in practice, but the CLT is!

The importance of the CLT is due to the speed of convergence:

$$P\left(\frac{\overline{X}-\mu}{\sigma/\sqrt{n}} \leq u\right) \approx P(Z \leq u)$$

for n "large enough". How large? It depends on the shape of the underlying distribution.

n	2	10	20	50
shape	Normal	Symmetric	Moderate Non-normal	Very non-normal

CLT example

Engine lifetimes follow a roughly symmetric distribution with mean 3.4 years and s.d. 2.1 years.

What is the chance that the average life of 25 engines exceeds 4 years?

$$P(\overline{X} > 4) = P\left(\frac{\overline{X} - 3.4}{2.1/\sqrt{25}} > \frac{4 - 3.4}{2.1/\sqrt{25}}\right)$$

$$\approx P(Z > 1.43) = 0.0766$$

An accurate plot for detecting deviation from Normal

Consider a sample of size 10 (say) from $N(0,1)$

On average such a sample will be centered at 0 with the same amount of probability between each point.

Normal Quantile Plot (NQP)

The Computer can calculate what these *theoretical* quantiles $z_{(1)}, z_{(2)}, \ldots, z_{(10)}$ should be.

$z_{(1)}$	1	2	3	4	5	6	7	8	9	10
Probabilities	0.06	0.16	0.26	0.35	0.45	0.55	0.65	0.74	0.84	0.94
Quantiles	-1.55	-1.00	-0.66	-0.38	-0.12	0.12	0.38	0.66	1.00	1.55

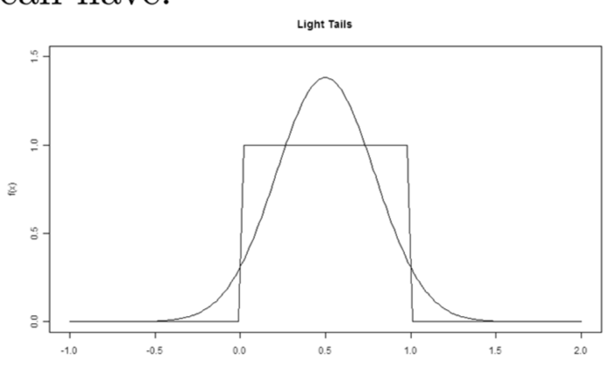
…and so on for any sample size n.

A *Normal Quantile Plot* of a univariate, numerical dataset, is a scatterplot of the data versus normal quantiles.

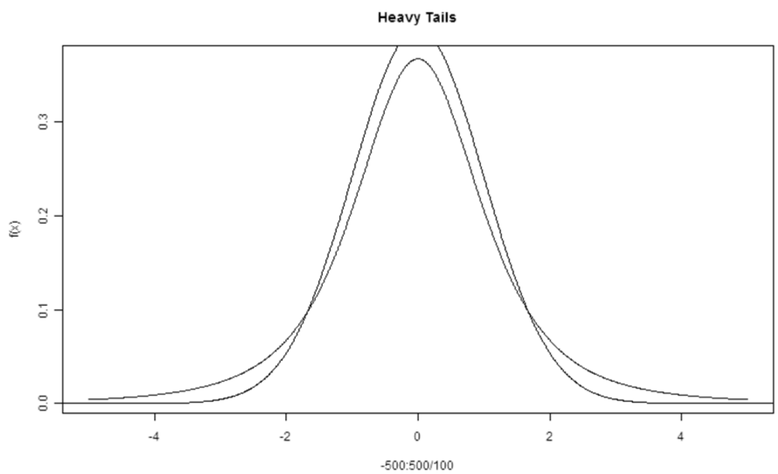
(Light and Heavy Tailed Distributions)

Right and left skew has already been defined. In addition, we can have:

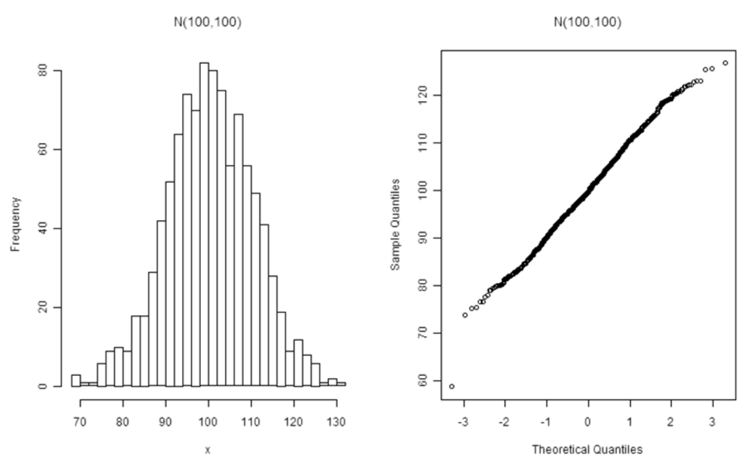
Light tails



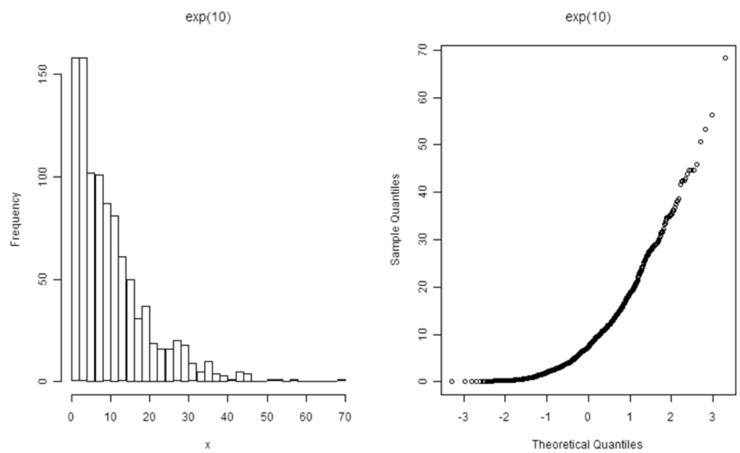
(Light and Heavy Tailed Distributions)



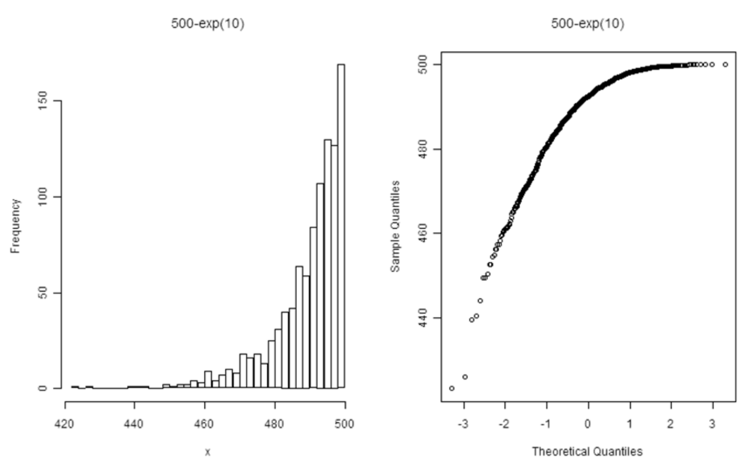
Interpretation: *Normal Data*



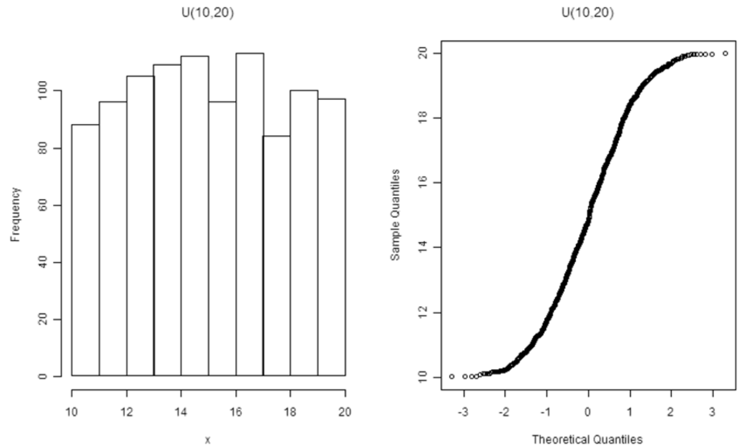
NQP Interpretation: *Right Skew*



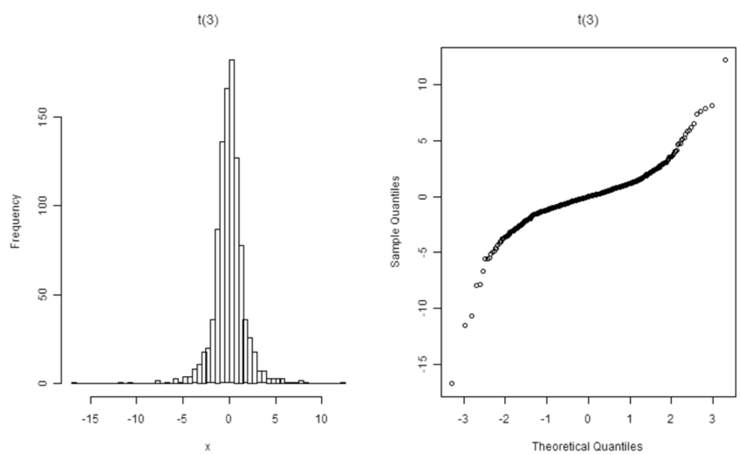
NQP Interpretation: *Left Skew*



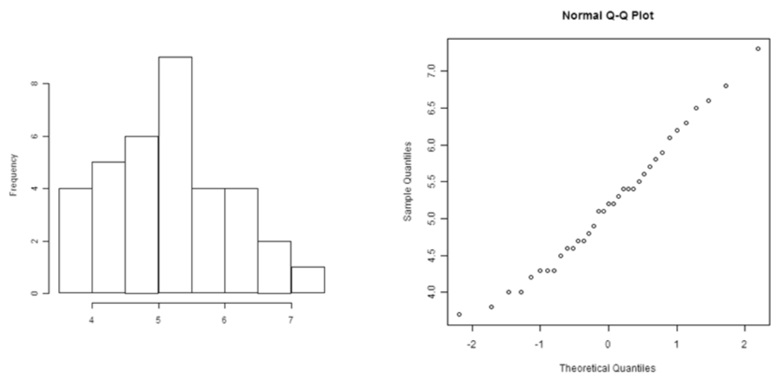
NQP Interpretation: *Light Tails*



NQP Interpretation: *Heavy Tails*



NQP also good for smaller samples



Mini-Recap – What are the Models in Use?

Wørd	Informal Definition	How Modeled?
Population	"All possible obversations under consideration"	Distribution X
Sample	"Subset of the population"	$X_1, \ldots, X_n$ Independent Same distribution
Dataset	$x_1, \ldots, x_n$ List of observations. A *realization of a sample.*	Nothing to model!

Mini-Recap: “Sample Average”

$$\frac{\overline{X}-\mu}{\sigma/\sqrt{n}}$$

Distribution if population is Normal

$N(0,1)$

If not →

Approximate distribution if sample size is “large enough”

$N(0,1)$

Verify using NQP

Definition – “Parameter”

A *parameter* (tends to be) a fixed constant that characterizes some aspect of the distrubution of a population. e.g.:

$$N(\mu, \sigma^2)$$

The parameters are μ and σ^2.

"Statistics" – The General Problem

The distribution used to model the population might not be fully specified.

$$N(4,2) \qquad N(\mu,2) \qquad N(\mu,\sigma^2)$$

"Some distribution with a mean and a variance (both unknown)"

$$Y_{ij} = \mu_i + \varepsilon_{ij}, \quad \varepsilon_{ij} \text{ have the same variance.}$$

Unknown quantities are often *parameters* of a distribution.

"Statistics" – The General Solution

Obtain a *sample*, e.g.: $X_1, \ldots, X_n$

(Or: $Y_{11}, \ldots, Y_{1n}, Y_{21}, \ldots, Y_{2n}, \ldots$, for example)

Use probability to *infer* some statement about the population. "Statistical inference"

In many cases it might be intuitively clear how to use the sample . . .

What types of inferences might be made?

Given a $N(\mu, \sigma^2)$ population, guess values for μ and/or σ^2: *Estimation.*

or

Given a model $Y_{ij} = \mu_i + \varepsilon_{ij}$, see if $\mu_1 = \cdots = \mu_I$, or not. *Hypothesis testing.*

or

Given a model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, predict the value of Y at a new input x. *Prediction.*

What types of inferences might be made?

A population has an unknown distribution. Determine what the distribution might be, from some suitable family of candidates.

Distribution fitting.

How will all this be done? Definition of *Statistic*

A *statistic* is a function of a sample.

Sample mean. $$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n}$$

Sample variance. $$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

In the $N(\mu, \sigma^2)$ case these *statistics* are used to make statements about the obvious corresponding parameters. CLT...

Agenda

- Overview of Statistical Inference
  - Estimation
  - Prediction
  - Hypothesis testing
- But there will be a digression…
  - …to discuss an important distribution related to the Normal distribution.

Point Estimation

Again: the population model X may not be completely specified:

$$N(\mu, 25) \qquad N(\mu, \sigma^2) \qquad F(\mu)$$

We can plan to obtain a sample: $X_1, X_2, \ldots, X_n$ to guess unknowns. The guesses will have certain properties.

The actual guesses themselves will use the observed sample $x_1, x_2, \ldots x_n$

Point "Estimator"

An *estimator* is a statistic (in particular, a random variable) used to guess the value of a parameter.

Desirable properties:

accurate (correct on average) $\longleftarrow E(\cdot)$

precise (not too variable)

consistent (improves as $n \nearrow$)

$\Big\}\longrightarrow \mathrm{Var}(\cdot)$

Example point estimators

Population: $N(\mu, 25)$ Sample $X_1, \ldots, X_n$

Usual estimator: $\overline{X}$ Stupid estimator: $\frac{X_1+X_2}{2}$

Both are "accurate":

$$E\left(\overline{X}\right) = \mu \qquad E\left(\frac{X_1 + X_2}{2}\right) = \mu$$

Example estimators

The usual estimator is more precise:

$$\mathrm{Var}\left(\overline{X}\right) = \frac{25}{n} \qquad \mathrm{Var}\left(\frac{X_1 + X_2}{2}\right) = \frac{25}{2}$$

Also from this, the usual estimator gets better with $n \nearrow$, but stupid one just sits there being stupid.

Most situations have a "usual" estimator.

Another example point estimator

One factor experiment with only two levels, 1 and 2. Population divided into two groups $N(\mu_1, 100)$ and $N(\mu_2, 100)$.

Usual question: what is the difference?

Estimate $\mu_1 - \mu_2$

Sample: $Y_{11}, \ldots, Y_{1n}, Y_{21}, \ldots Y_{2n}$

Usual estimator: $\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}$

Another example estimator

$$E\left(\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}\right) = \mu_1 - \mu_2$$

$$\begin{aligned}\mathrm{Var}\left(\overline{Y}_{1\cdot} - \overline{Y}_{2\cdot}\right) &= \mathrm{Var}\left(\overline{Y}_{1\cdot}\right) + \mathrm{Var}\left(\overline{Y}_{2\cdot}\right)\\ &= \frac{100+100}{n}\end{aligned}$$

Interval Estimation

Point estimators do not provide any assessment of precision.

One could instead report a range of plausible values based on two statistics L and U that satisfy (e.g. in the case of estimating a mean μ):

$$P(L \leq \mu \leq U) = 1 - \alpha$$

“Confidence Interval”

$$P(L \leq \mu \leq U) = 1 - \alpha$$

(L, U) is a $100(1-\alpha)\%$ *confidence interval*

L and U are the *confidence limits*.

$100(1-\alpha)$ is the *confidence level* and is typically close to 100, such as 90, 95, 99 etc.

Example of such an L and U

Population: $N(\mu, 25)$ Sample $X_1, \ldots, X_n$

$$\alpha = 0.05$$

$$P\left(-1.96 \leq \frac{\overline{X} - \mu}{5/\sqrt{n}} \leq 1.96\right) = 0.95$$

$$P\left(\overline{X} - 1.96\frac{5}{\sqrt{n}} \leq \mu \leq \overline{X} + 1.96\frac{5}{\sqrt{n}}\right) = 0.95$$

Informally: "$\overline{X} \pm 1.96\frac{5}{\sqrt{n}}$"

Numerical Example

Calcium concentration in an oil additive is $N(\mu, 400)$. A sample of size 25 is taken. The observed sample average is 491ppm.

A 95% confidence interval for μ is:

$$491 \pm 1.96\frac{20}{5} = (483.16, 498.84)$$

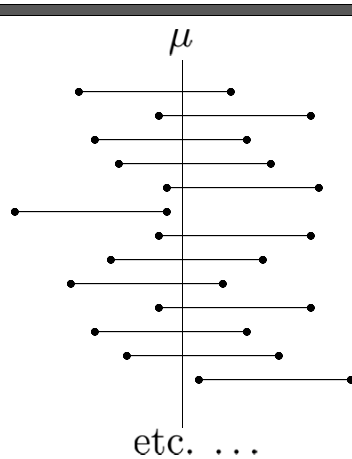
Use/Myth/Meaning

USE: To give a range of plausible values for μ that accounts for the sampling procedure undertaken.

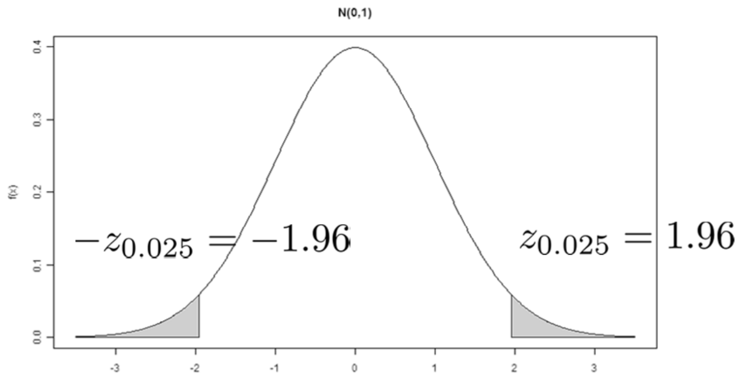
MYTH: $P(483.16 \leq \mu \leq 498.84) = 0.95$

LITERAL MEANING: If many (hypothetical) samples were observed, and a C.I. computed each time, then $100(1-\alpha)\%$ of the intervals will contain μ.

Meaning



What influences the width of a confidence interval?



Generally: $\overline{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$

Sample size for a given "margin of error"

The difference between estimator and parameter (e.g. $\overline{X}$ and μ) can be called *margin of error* denoted by m

Problem: determine n so that $\left|\overline{X} - \mu\right| < m$ with probability $1 - \alpha$.

Solution: since $|m| = \left|z_{\alpha/2} \frac{\sigma}{\sqrt{n}}\right|$, just solve for n in:

$$m = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \Longrightarrow n = \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2$$

Sample size numerical example

Calcium concentration in an oil additive is $N(\mu, 400)$. What sample size is required to obtain an estimate for μ within 5ppm with probability 0.95?

$$n = \left(\frac{z_{\alpha/2}\sigma}{m}\right)^2 = \left(\frac{1.96 \cdot 20}{5}\right)^2$$
$$= 61.4656$$

Sample size required is 62 (conservative).

A more realistic situation

It is unlikely to have $N(\mu, 400)$, say, as a population model, i.e. with variance known.

More realistic: $N(\mu, \sigma^2)$ both parameters unknown.

Confidence interval based on:

$$P\left(-z_{\alpha/2} \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq z_{\alpha/2}\right) = 1 - \alpha$$

What to do about the unknown variance?

What is the obvious thing to try?

Estimate σ^2 with something—obviously (?) with the sample variance s^2

Proposed new formula: $\overline{X} \pm z_{\alpha/2} \frac{s}{\sqrt{n}}$

Is it valid? No - the "coverage" probability is too low for small to moderate n.

A new (family of) distribution(s)

Sample $X_1, X_2, \ldots, X_n$ from a $N(\mu, \sigma^2)$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \qquad \frac{\overline{X} - \mu}{s/\sqrt{n}}$$

"t distribution" with parameter $n-1$, where n is the sample size.

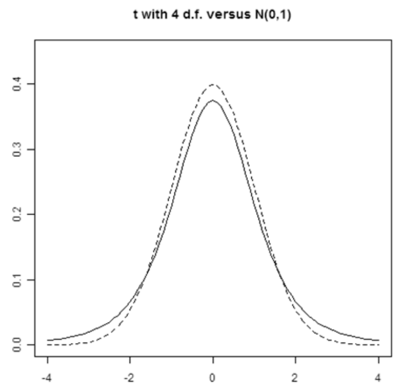
Notation: t_{n-1}

Fundamental Difference between *t* and *Z*

Densities:

$$f_Z(z) \propto e^{-z^2/2}$$

$$f_t(t) \propto \left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}$$



"One Sample *t* confidence interval for the mean"

$$\overline{X} \pm t_{n-1,\alpha/2}\frac{s}{\sqrt{n}}$$

Calcium concentration in an oil additive is $N(\mu, \sigma^2)$. A sample of size 25 is taken. The observed sample average is 491ppm and the observed sample variance is 426..

A 95% confidence interval for μ is:

$$491 \pm 2.064\frac{\sqrt{426}}{5} = (482.47, 499.56)$$