MIE 1807
Principles of Measurement

2017-01-30
Neil Montgomery

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

Some Rules

$$E(aX + b) = aE(X) + b$$

$$E(X_1 + X_2) = E(X_1) + E(X_2)$$

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

If X_1 and X_2 are independent:

$$\mathrm{Var}(X_1 + X_2) = \mathrm{Var}(X_1) + \mathrm{Var}(X_2)$$

but not vice-versa.

Standardization

The rules allow us to consider *standardized* versions of random variables, with mean 0 and variance 1.

$$Y = \frac{X - \mu}{\sigma}$$

$$E(Y) = E\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma}(\mu - \mu) = 0$$

$$\mathrm{Var}(Y) = \mathrm{Var}\left(\frac{X - \mu}{\sigma}\right) = \frac{1}{\sigma^2}\sigma^2 = 1$$

But Y and X have the same *shape*.

A frequent probabilistic situation…

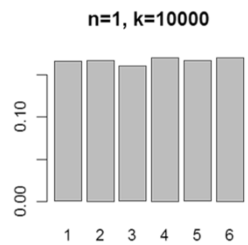
- Many random quantities result from:
  - Sums/averages of…
  - …a relatively large number of…
  - …roughly equally weighted…
  - …independent random contributions.
- Direct examples: height, test scores, measurement errors.
- Indirect example: *sample averages*

Indirect Example 1

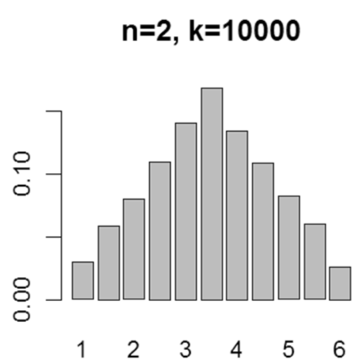
Toss n dice and let X be the sample average result. Formally:

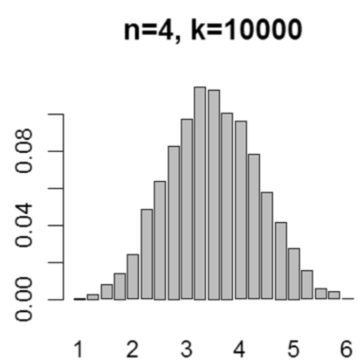
$$X = (X_1 + X_2 + \cdots + X_n)/n$$

Histogram of $k = 10000$ tosses with $n = 1$

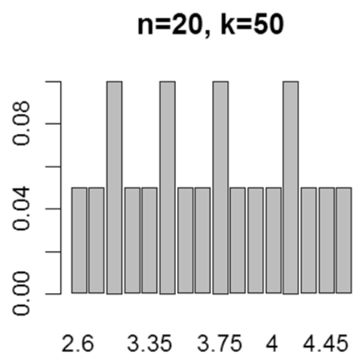


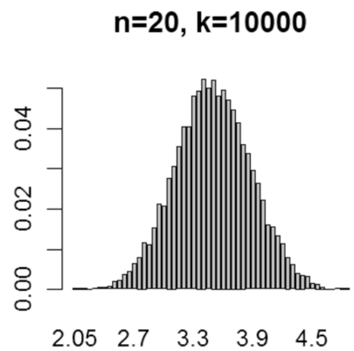
Indirect Example 1



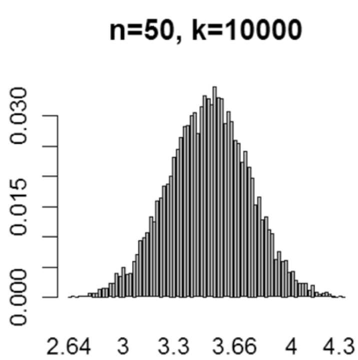


Indirect Example 1





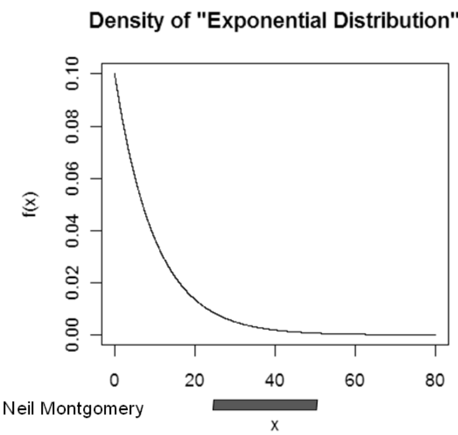
Indirect Example 1



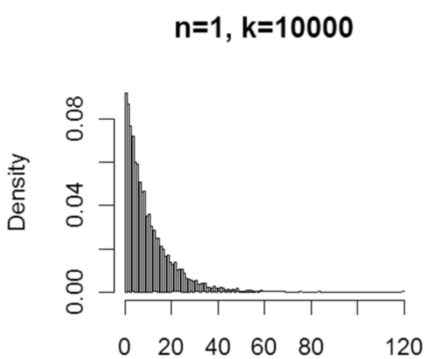
Indirect Example 2

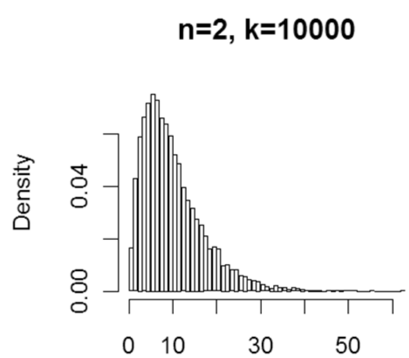
Observe n engine failures X be the sample average failure time.

Density commonly used to model:
$f(x) = \lambda e^{-\lambda x}$ for $x > 0$

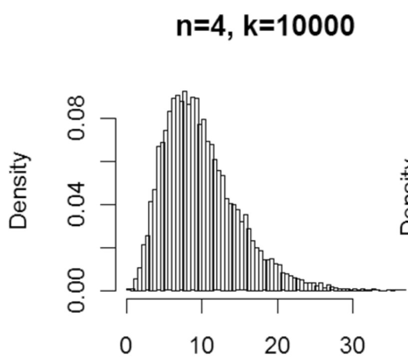


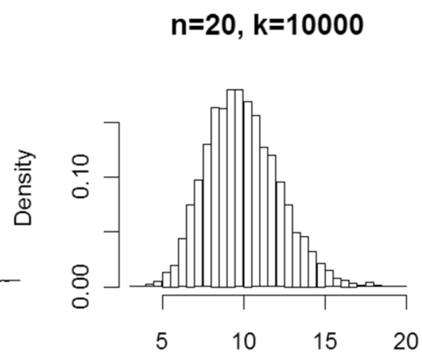
Indirect Example 2



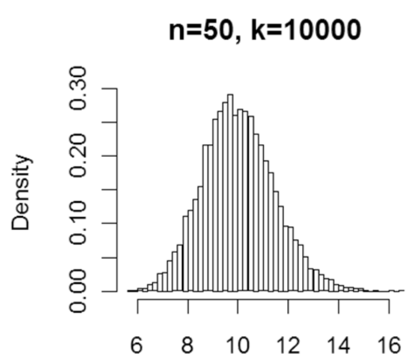


Indirect Example 2





Indirect Example 2



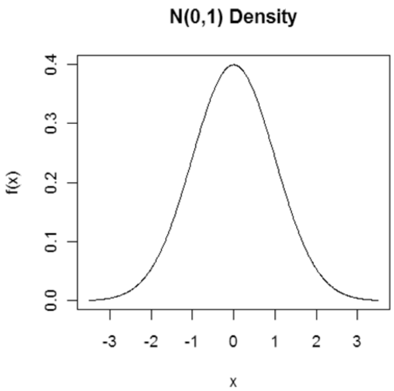
Model: The *Normal* Distributions

Symmetric, bell-shaped densities:

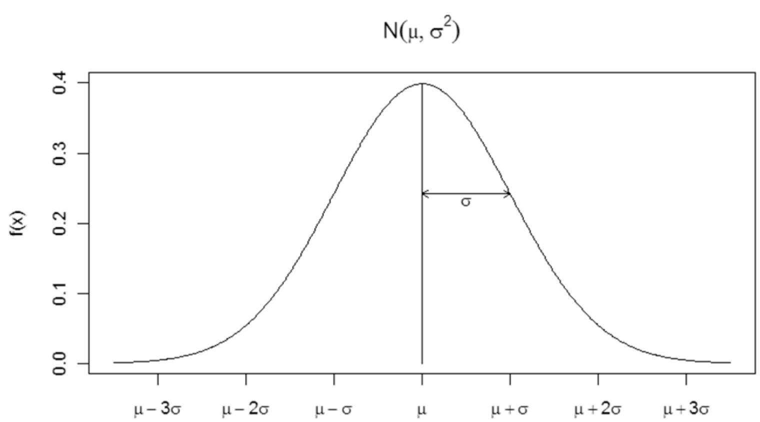
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$E(X) = \mu$$

$$\mathrm{Var}(X) = \sigma^2$$



Normal densities all have the same shape



The Normal Distributions

"X has a Normal distribution with mean μ and variance σ^2:

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

"Standard"(ized) normal denoted (conventionally) by Z, and is such that

$$Z \sim N(0,1)$$

Relationship:

$$Z = \frac{X-\mu}{\sigma} \qquad X = \mu + \sigma Z$$

Normal Probability Computations

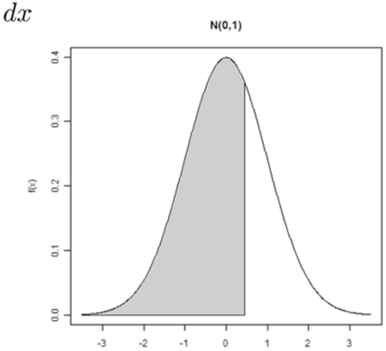
If $X \sim N(\mu, \sigma^2)$ then

$$P(a < X < b) = \int_a^b \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \, dx$$

Explicit antiderivative not available.

Real world: use computer.

Tests: transform $Z = \frac{X-\mu}{\sigma}$ and use table of standard Normal probabilities.



Interesting fact: Normal + Normal ~ Normal

If $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$, then $X_1 + X_2 \ldots$ is a random variable too.

In fact $X_1 + X_2$ also has a Normal distribution.

If $X_1 \perp X_2$:

$$X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

Model for notion of "Sample"

We want a mathematical model for "replication" (of an experiment).

In other words, we got the numbers $x_1, x_2, \ldots, x_n$ in our dataset, but where did they come from?

Define a *sample* as a sequence of independent r.v.s with the same distribution: $X_1, X_2, \ldots, X_n$

"i.i.d."

Interesting Functions of Samples

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

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

These are random variables.

(Sorry for even more confusion!)

Given a dataset $x_1, \ldots, x_n$

$$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad s^2 = \frac{\sum_{i=1}^{n}(x_i - \overline{x})^2}{n-1}$$

were *also* called sample mean and sample variance, but these are just numbers!

Call these quantities *observed*...

Mean and Variance of "Sample Average"

For *any sample* $X_1, \dots, X_n$ that are independent and have the same distribution with mean μ and variance σ^2:

$$E(\overline{X}) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n} n\mu = \mu$$

$$\mathrm{Var}(\overline{X}) = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$$

$$SD(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

Special Case – Normal

Suppose further that $X_1, \dots, X_n$ is a sample from a $N(\mu, \sigma^2)$ distribution.

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \qquad \overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

“Fundamental Theorem of Statistics”

For *any* population modeled with a distribution with mean μ and variance σ^2.

Consider a sample $X_1, \ldots, X_n$.

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

Where $Z \sim N(0,1)$.

Really called the “Central Limit Theorem” or CLT.

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.

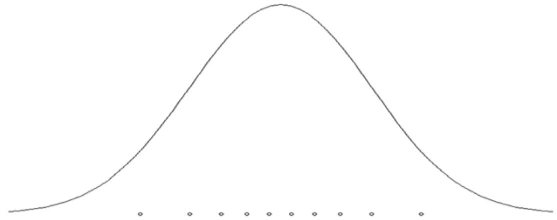
What is the chance that the average life of 25 engines exceeds 4 years?

$$P(\overline{X} > 4) = P\left(\frac{\overline{X} - 3.4}{2.1/\sqrt{25}} > \frac{4 - 3.4}{2.1/\sqrt{25}}\right)$$
$$\approx P(Z > 1.43) = 0.0766$$

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

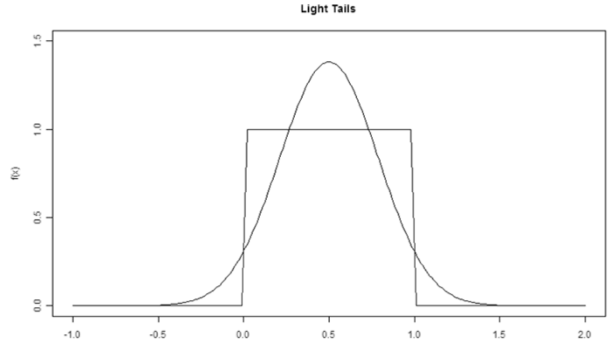
. . . and so on for any sample size n.

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

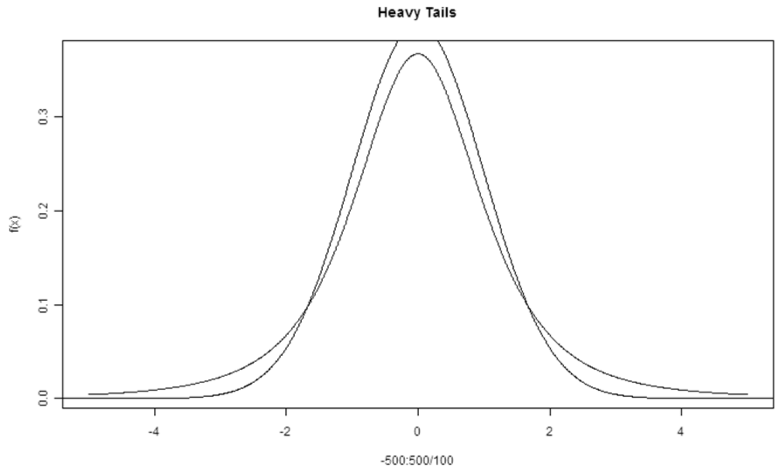
(Light and Heavy Tailed Distributions)

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

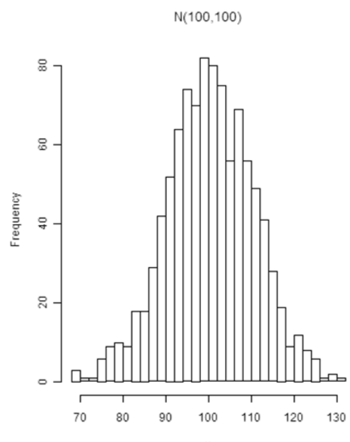
Light tails

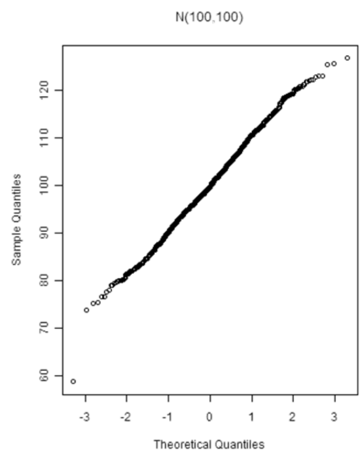


(Light and Heavy Tailed Distributions)

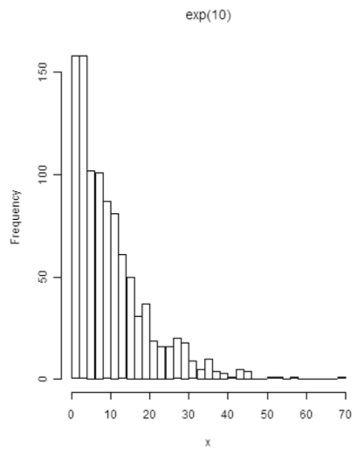


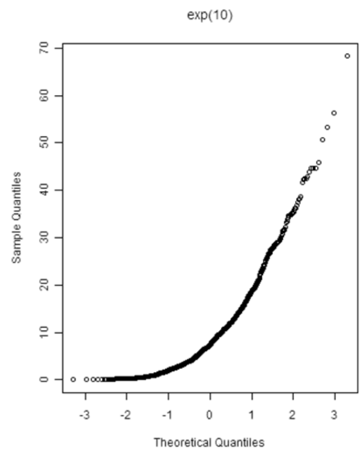
Interpretation: *Normal Data*



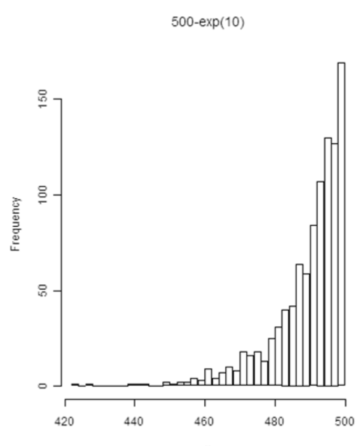


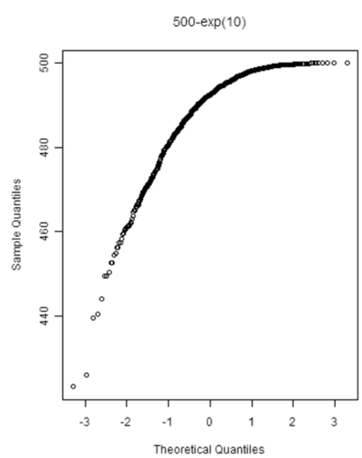
NQP Interpretation: *Right Skew*



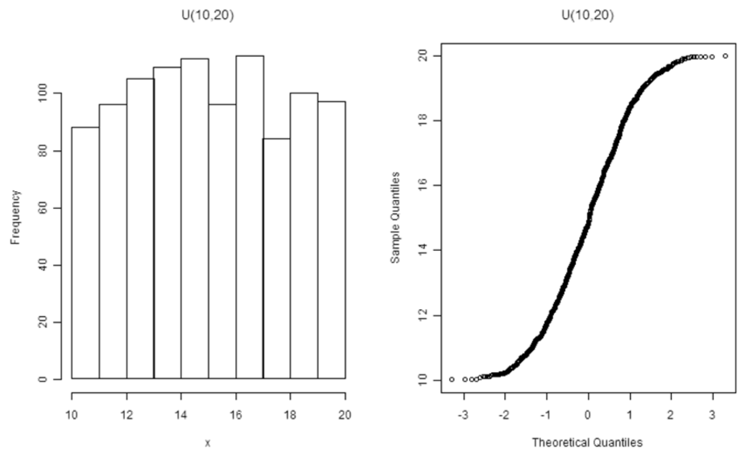


NQP Interpretation: *Left Skew*

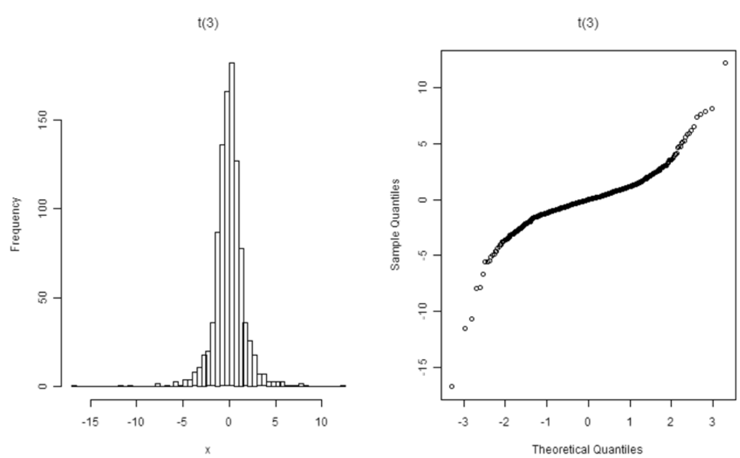




NQP Interpretation: *Light Tails*



NQP Interpretation: *Heavy Tails*



NQP also good for smaller samples