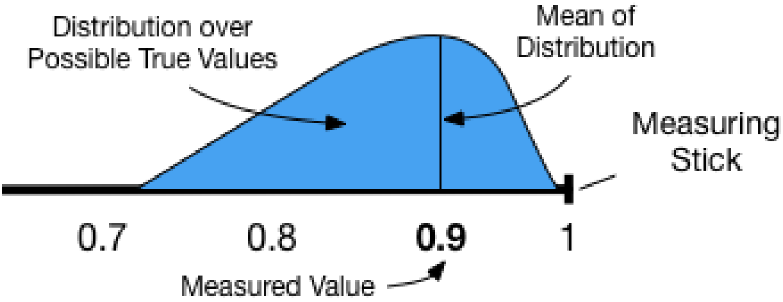
Simulation and Estimation

Justin Silverman, MD, PhD

The Pennsylvania State University
College of Information Science and Technology
Departments of Statistics and Medicine
Institute for Computational and Data Science

Uncertainty Propagation

All Measurements Have Error



Often assume error is symmetric and normally distributed about measured value – but it need not be.

The Problem of Uncertainty Propagation

Suppose we measured $\hat{x}$ which we assume is subject to normally distributed error such that the true value can be described by distribution $x \sim N(\hat{x}, 1)$.

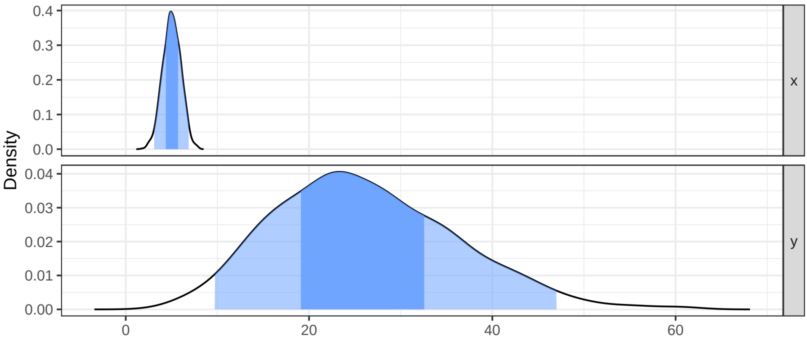
Problem Statement

Suppose we want to use x in a calculation, e.g., $y = f(x)$. How does error in x propagate into error in y?

Example

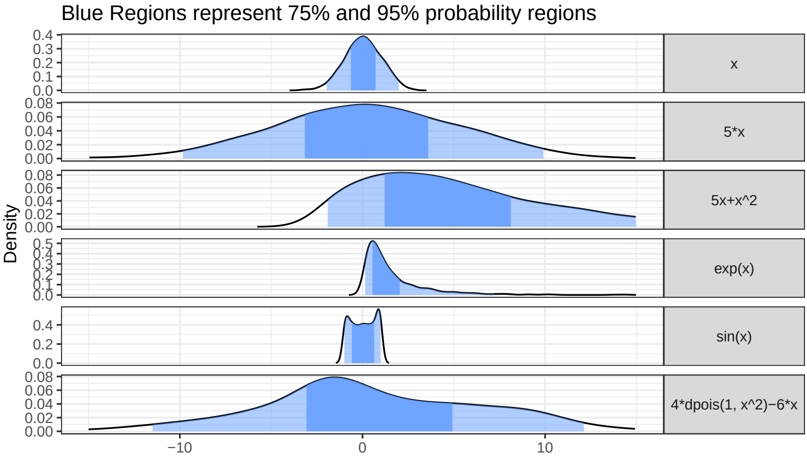
Suppose $\hat{x} = 5$ and we want to calculate $y = x^2$. What is the distribution of y?

```r
x <- rnorm(n=1000, mean=5, sd=1)
y <- x^2
```



More Examples

Suppose $x \sim N(0, 1)$ for simplicity.

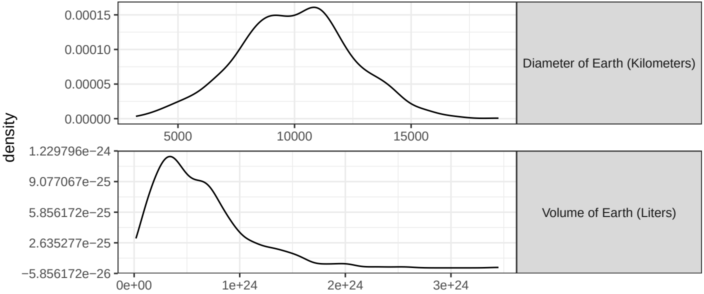


Improved Back of the Envelope Calculations

Whats the volume of the earth?

- I guessed its got a diameter of about 10,000 kilometers
- Assume my error is normally distributed, I guess a standard deviation of 2500km
  - 68% chance true value is between 7,500 and 12,500 kilometers.

```r
t <- 1000
kms <- rnorm(t, mean=10000, sd=2500)  # sample values our guessed distribution
volume.km3 <- (4/3)*pi*(kms/2)^3      # volume in cubic kilometers
liters <- volume.km3*1e12             # convert to liters
```



The true answer is 1.083e+24 (86th percentile of my guess).

Estimation and Hypothesis Testing

Estimation as Uncertainty Propagation

In (Frequentist) statistics, the sampling distribution of an estimator is a type of error propagation.

Definition (Estimator)

Any function (f) which we apply to data (x) to estimate something (μ). *An Estimator is a statistic.*

The sample mean is an estimator of the population mean:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N} x_n.$$

Sampling Distribution of an Estimator

Assume that our data is *iid* from an underlying population $x_1, \ldots, x_N \sim N(\mu, \sigma^2)$.

What is the distribution of the sample mean? (Uncertainty Propagation)

This distribution is called the sampling distribution of the estimator.

From Sampling Distributions to Hypothesis Testing

Assume $x_1, \ldots, x_N \sim N(\mu, \sigma^2)$. We want to know the sampling distribution of the sample mean:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n.$$

Why? (many reasons, will just pick one)

Common Statistics Logic: Use Data To Reject a (Null) Hypothesis

Suppose $H_0 : \mu = 0$, then $x_1, \ldots, x_N \sim N(0, \sigma^2)$.

Then calculate the sampling distribution of $\hat{\mu}$ and see if the value we observe is unlikely (e.g., $<$0.05% probability) under this null model. That probability value is a p-value of a hypothesis test.

But we don't know σ^2 so we can't calculate the sampling distribution of $\hat{\mu}$!

Building the t-test

Assume (under H_0, Null Hypothesis) that $x_1, \ldots, x_N \sim N(0, \sigma^2)$. We want to know the sampling distribution of the sample mean:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

but we can't calculate that because we don't know σ^2.

Building the t-test

Assume (under H_0, Null Hypothesis) that $x_1, \ldots, x_N \sim N(0, \sigma^2)$. We want to know the sampling distribution of the sample mean:

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n$$

but we can't calculate that because we don't know σ^2.
(Statisticians) **Let's change the statistic:**

$$t = \sqrt{N}\hat{\mu}/\hat{\sigma}$$

where $\hat{\sigma}$ is itself an unbiased estimator of the population standard deviation

$$\hat{\sigma} = \sqrt{\frac{\sum_{n=1}^{N}(x - \hat{\mu})^2}{n-1}}.$$

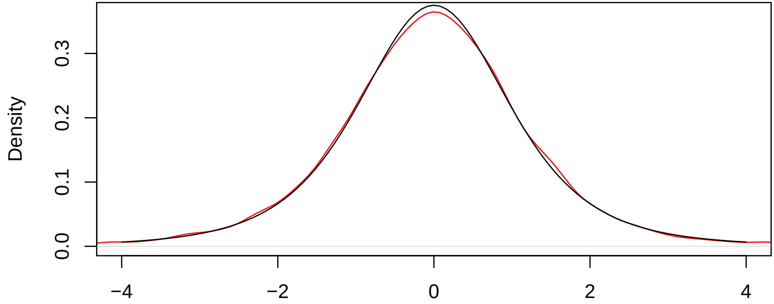
Some clever math shows that this new statistic t does not depend on any unknowns of the population (its a *pivotal quantity*):

$$t \sim t_{n-1}(\text{Student's t with } n-1 \text{ degrees of freedom}).$$

Checking Our Work 1

```
N <- 5
n.sim <- 10000
t <- rep(NA, n.sim)
for (i in 1:n.sim) {
  x <- rnorm(N, mean=0, sd=2)
  t[i] <- sqrt(N) * mean(x)/sd(x)
}
```

Plot (red) and compare to t_{N-1} density (black)



Checking Our Work 2

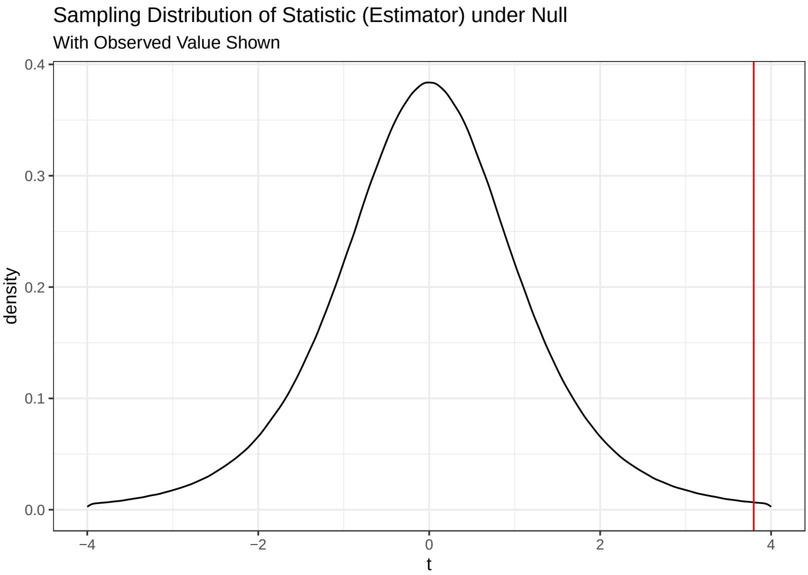
```
x <- c(3,4,3,3,1,0)

N <- length(x)
s <- sd(x) # same as sqrt(sum((x-mean(x))^2)/(N-1))
t <- sqrt(N)*mean(x)/s

## simulate sampling distribution of t (under null)
## and compare observed value
t.sim <- rt(n=10000000, df=N-1)

## calculate the number of simulations where |t|<|t.sim|
## this is the p-value
## this is the probability of the observed statistic under the null  model
sum(abs(t)<abs(t.sim))/length(t.sim)
```

```
[1] 0.0127504
```

```
## now compare to closed-form (known) solution
t.test(x)$p.value
```

```
[1] 0.0126766
```

Visualization Based on Simulation



Big Picture

(Frequentist) estimation and hypothesis testing are just clever applications of uncertainty quantification.

Differences:

- In uncertainty quantification, we typically think of randomness as arising due to measurement error.
- In estimation and hypothesis testing, we typically think of randomness as arising due to random sampling of a larger population.

Thinking of this in terms of simulation also makes it easy to do power and sample-size calculations.

Power and Sample Size Calculations

Power and Sample Size Calculation

- Up until now we have been operating under the null (sampling distribution of statistic under null).
- Now we focus on the alternative and ask one of two questions.

Power Calculations

Given my test and my belief in what the true population distribution is: for a given sample size, what is the probably that I reject the null hypothesis (the power of the test)?

Sample Size Calculations

Given my test and my belief in what the true population distribution is: what sample size do I need to achieve a desired power?

Power Calculations Made Easy (Example)

- For convenience lets use the t-test we already discussed and test the null hypothesis $H_0 : \mu = 0$.
- Suppose I believe $H_A : \mu = 5$ and $\sigma^2 = 3$.

```
n.sim <- 1000
N <- seq(3, 20)
power <- rep(NA, length(N))
for (n in 1:length(N)) {
  reject <- rep(NA, n.sim)
  for (i in 1:n.sim) {
    x <- rnorm(N[n], mean=5, sd=3)
    p.val <- t.test(x)$p.value
    reject[i] <- p.val <= 0.05
  }
  power[n] <- sum(reject)/n.sim
}
plot(N, power, type="l")
```