Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

February 22, 2016

The Voinovich School of Leadership and Public Affairs

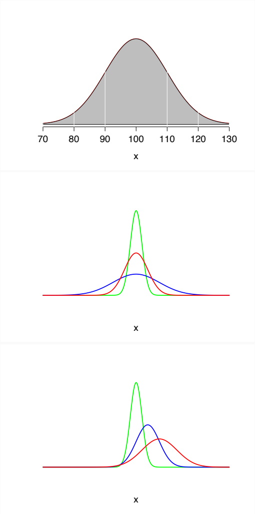
Table of Contents

1. The Normal Distribution
2. The Standard Normal Distribution
3. The Central Limit Theorem
4. $\sim N()$ Approximation for Binomial

The Normal Distribution

$\sim N()$

1. The pdf of $\sim N()$:
   $f(Y) = \dfrac{1}{\sigma\sqrt{2\pi}} e^{-(Y-\mu)^2/2\sigma^2}$
2. $\mu =$ mean; $\sigma =$ s.d., $\pi = 3.14159$, and; $e = 2.71828$
3. Two parameters $(\mu;\sigma)$ describe a normal distribution
4. The normal distribution is symmetric, with the tails extending from $-\infty$ to $+\infty$
5. μ can be positive, negative, or zero
6. σ dictates how wide or flat the curve is
7. The total area under the normal curve is 1, and the area to the left of μ is 0.5; the area to the right of μ is 0.5



Standardizing Variables: The $z - score$

Definition

Z-scores allow us to identify the relative location of an observation in a data set by telling us how many standard deviation units above or below the Mean a particular value (Y_i) falls

- The z-score, also known as the standardized value, is $z_i = \frac{Y_i - \bar{Y}}{s}$
- The z-score allows us to compare scores drawn from distributions with dissimilar variability (see below)
- z-scores greater/smaller than ± 3 are indicative of outliers

Place	$(\mu; \sigma)$	December	z
Caribou (ME)	(110 inches; 30 inches)	125 inches	+0.50
Boston (MA)	(24 inches; 5 inches)	39 inches	+1.50

- Which place had more unusual snowfall last month?

Chebyshev's Theorem & The Empirical Rule

Definition

Chebyshev's Theorem: At least $(1 - \frac{1}{z^2})$ of the data values must lie within z standard deviations of the mean, where z is any value greater than 1

z	$(1 - \frac{1}{z^2})$	Interval
2	$(1 - \frac{1}{2^2})$	0.75 (75%)
3	$(1 - \frac{1}{3^2})$	0.89 (89%)
4	$(1 - \frac{1}{4^2})$	0.94 (94%)

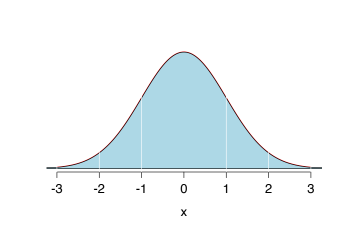
- The Empirical Rule extends Chebyshev's Theorem to ***bell-shaped distributions***
    1. About 68% of the data values fall within $\pm$ 1 standard deviation
    2. About 95% of the data values fall within $\pm$ 2 standard deviation
    3. About 99% of the data values fall within $\pm$ 3 standard deviation

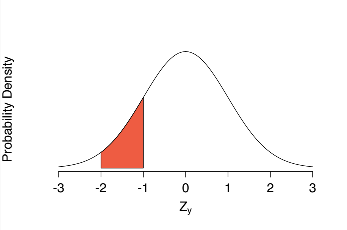
The Standard Normal Distribution

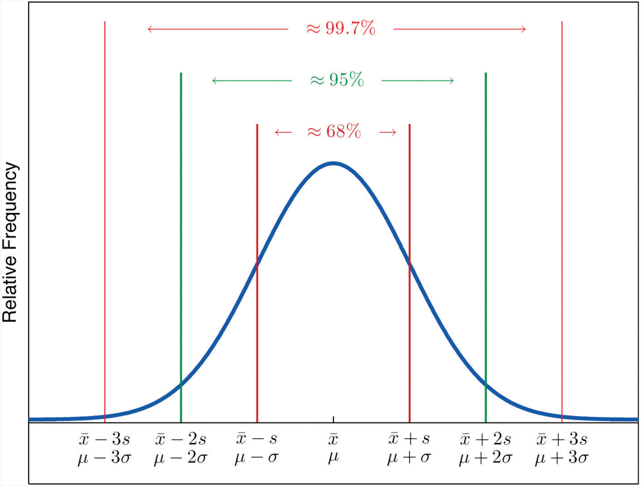
Standardizing Variables: The $z-score$

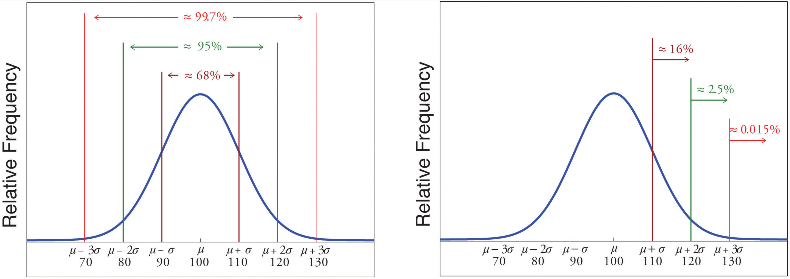
Z-scores allow us to identify the relative location of an observation in a data set by telling us how many standard deviation units above or below the mean (μ) a particular value (Y_i) falls

- $z_i = \frac{Y_i - \bar{Y}}{s}$
- The standard normal distribution $\sim N(\mu = 0; \sigma = 1)$
- 68% of data within $\pm$ 1 standard deviations
- 95% of data within $\pm$ 2 standard deviations
- 99% of data within $\pm$ 3 standard deviations
- z-scores greater/smaller than ± 3 are indicative of outliers






≈ 99.7%
≈ 95%
≈ 68%
Relative Frequency
$\bar{x} - 3s$ $\mu - 3\sigma$
$\bar{x} - 2s$ $\mu - 2\sigma$
$\bar{x} - s$ $\mu - \sigma$
$\bar{x}$ μ
$\bar{x} + s$ $\mu + \sigma$
$\bar{x} + 2s$ $\mu + 2\sigma$
$\bar{x} + 3s$ $\mu + 3\sigma$

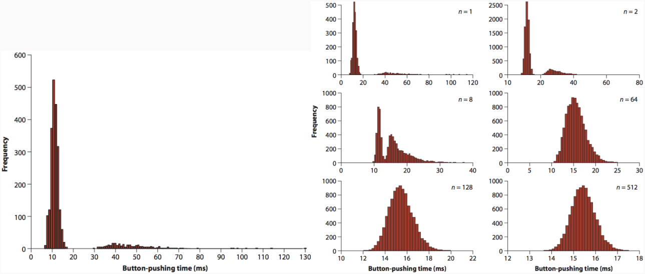


(a) Whole Spectrum

(b) Higher End

The Central Limit Theorem

The Central Limit Theorem



Definition

The sampling distribution of $\bar{Y}$ with random sample size n will be normally distributed with mean $= \mu$ and standard deviation $= \dfrac{\sigma}{\sqrt{n}}$ if

1. the population is normally distributed or
2. the sample size $n \geq 30$

$\sim N()$ Approximation for Binomial

$\sim N()$ Approximation for the Binomial Distribution

What the Central Limit Theorem also tells us is that the sum (and mean) of independently and identically distributed random variables is itself approximately normally distributed if the sample size is large enough $(n \geq 30)$

For problems involving the Binomial distribution but involving large samples we could use the Normal approximation

With large n the binomial distribution for the number of successes is approximately Normally distributed with mean $= np$ and standard deviation $= \sqrt{np(1-p)}$ so long as:

1. $np > 5$, and
2. $n(1-p) > 5$

The Brown Recluse Spider Example

H_0: These spiders have no preference for live/dead prey ($p = 0.50$)
H_A: These spiders have a preference either for live/dead prey ($p \neq 0.50$)
In a study 31 out of 41 spiders ate dead prey
Using the Binomial approach we would test $P(31\text{ Successes})$ if H_0 were true
This would yield $P - value = 0.00145$ and so we would Reject H_0

Can we use the Normal approximation? Let us check:

1. $np = 41 \times 0.50 = 20.5$
2. $n(1-p) = 41(1-0.50) = 41 \times 0.50 = 20.5$

Now, $\mu = np = 20.5$ and standard deviation
$\sigma = \sqrt{np\,(1-p)} = \sqrt{20.5\,(0.50)} = 3.2015$. So what is $P\,(X \geq 31)$?

1. Convert $X = 31$ into its $z - score$ via $z = \dfrac{X-\mu}{\sigma} = \dfrac{31-20.5}{3.2015} = 3.28$
2. What is the probability of getting a z-score this far in either tail (because it is a two-tailed test) by chance if H_0 is true? R tells us the $P - value = 0.001$ and so we Reject H_0; it is very likely that spiders show a dietary preference for dead prey.