Statistical Methods for Plant Biology

PBIO 3150/5150

Anirudh V. S. Ruhil

January 26, 2016

The Voinovich School of Leadership and Public Affairs

Table of Contents

1 Sampling Distributions

2 Measuring Uncertainty around an Estimate

The Standard Error of an Estimate

Confidence Intervals

3 Worked Examples

Sampling Distributions

Sampling Distributions

- Recall population values are parameters ... μ, σ^2, σ ... while our sample values are estimates ... $\bar{Y}, s^2, s$
- In fact, these sample values are point estimates ... single values that are supposed to reflect their corresponding population parameters

Definition

A point estimator is a sample statistic that predicts the value of the corresponding population parameter

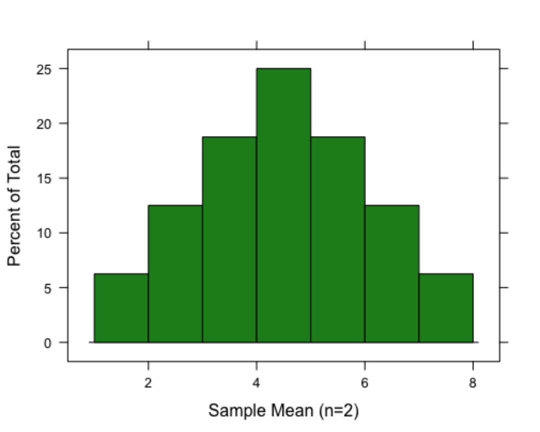
- Desirable point estimators have the following properties ...
  1. Sampling distribution of the point estimator is centered around the population parameter (*unbiasedness*)
  2. Point estimator has the smallest possible standard deviation (*efficiency*)
  3. Point estimator tends toward the population parameter as the sample size increases (*consistency*)
- What guarantees that these hold? Let us see ...

Understanding Sampling Distributions

Let a *population* of four scores be $[2,4,6,8]$. How many random samples of two scores can we construct, and what would the sample mean be in each sample? *Note:* $N = 4; n = 2$

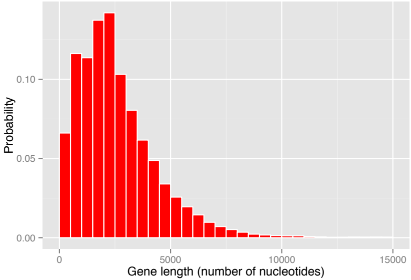
#	Y_1	Y_2	$\bar{Y}$	#	Y_1	Y_2	$\bar{Y}$
1	2	2	2	9	6	2	4
2	2	4	3	10	6	4	5
3	2	6	4	11	6	6	6
4	2	8	5	12	6	8	7
5	4	2	3	13	8	2	5
6	4	4	4	14	8	4	6
7	4	6	5	15	8	6	7
8	4	8	6	16	8	8	8

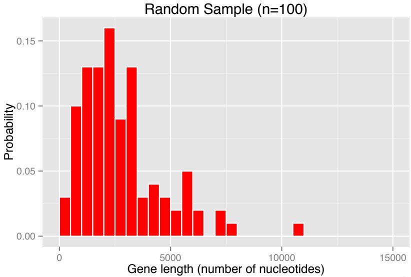
Plotting the Distribution of Sample Means



Mapping the Genome Population

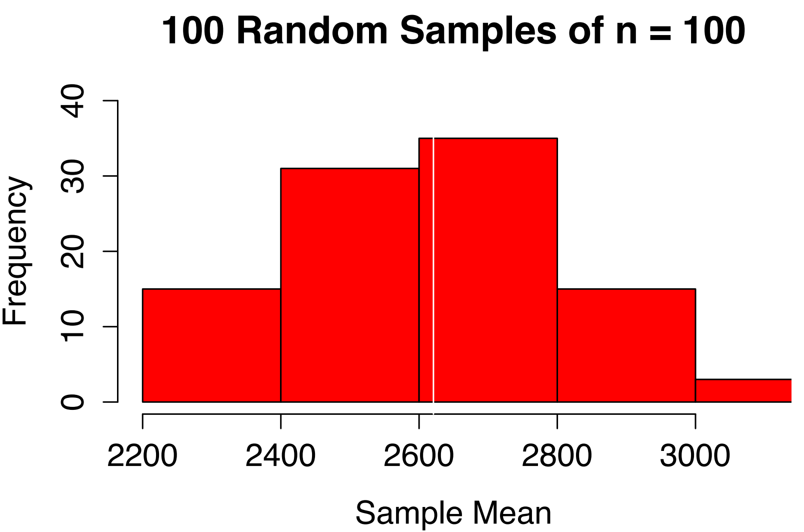
- Human Genome Project identified approximately 20,500 genes in human beings
- Top panel: Population of gene lengths ($N = 20,290$)
- Parameters: $\mu = 2,622$; $\sigma = 2,036.967$; $Min = 60$; $Max = 99,631$
- Bottom panel: Random sample of gene lengths ($n = 100$)
- Estimates: $\bar{Y} = 2,777$; $s = 1,875.814$; $Min = 87$; $Max = 10,503$





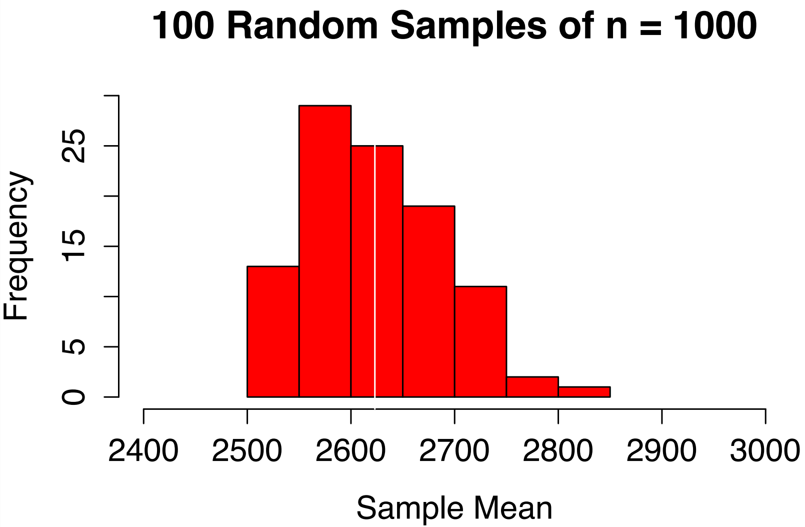
What if we drew multiple samples?

$\mu = 2622$



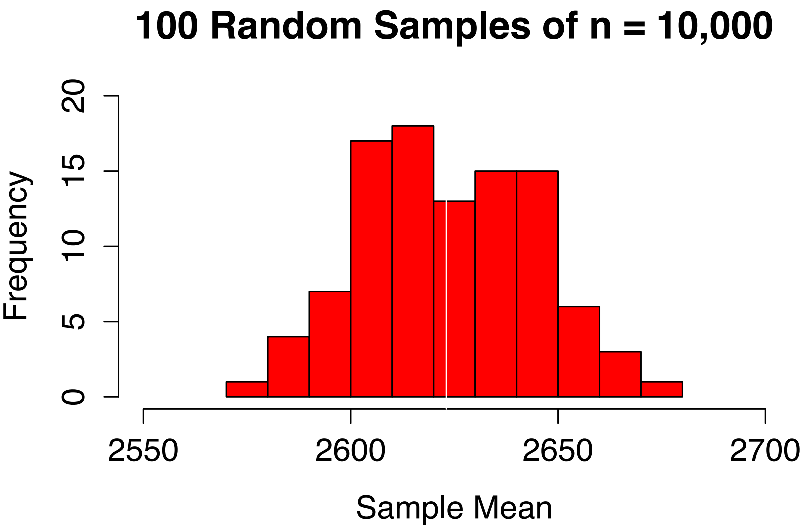
But what if we increased the sample size for each draw?

$\mu = 2622$



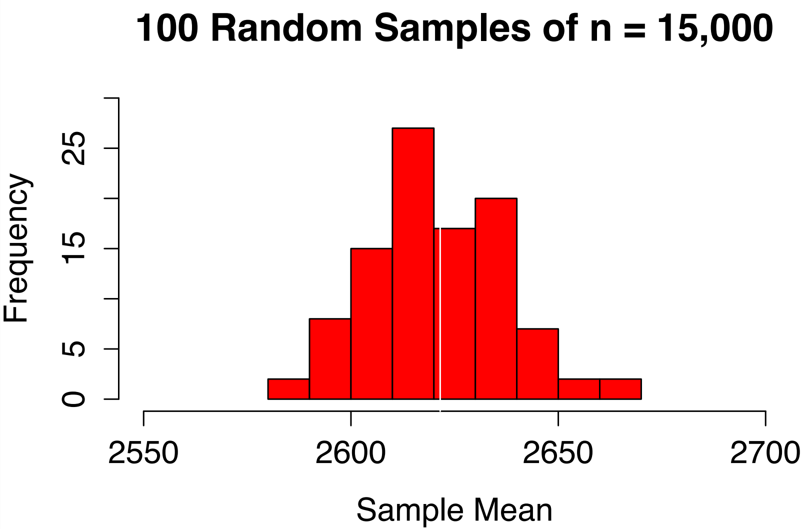
What if we increased the sample size even further?

$\mu = 2622$



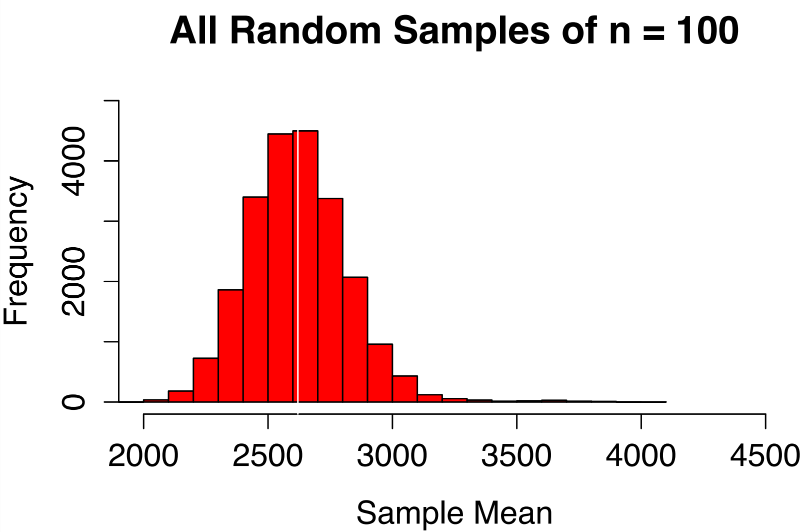
What if we increased the sample size even further?

$\mu = 2622$



What if we drew all possible samples of n = 100?

$\mu = 2622$



The Sampling Distribution

Definition

The sampling distribution of $\bar{Y}$ is the probability distribution of all possible values of the sample mean $\bar{Y}$

- What we are saying is that for any given random sample the expected value of $\bar{Y}$, denoted as $E(\bar{Y})$, $= \mu$
- Intuitively, unless we mess up our sampling, on average we should end up with a sample mean that equals the population mean (because the population mean has the highest frequency of occurrence in the population)
- The preceding simulations show that the larger the sample, the more likely we are to end up with a sample mean close to the population mean … larger samples yield more precise estimates
- “Likely to equal the μ” is one thing but how can we measure the precision of our sample-based estimate of the population mean?

Measuring Uncertainty around an Estimate

Measuring Uncertainty around an Estimate

- The question now is: How far would we expect, on average, our sample mean to be from the population mean, for a given sample size?
- The standard error provides the answer: $\sigma_{\bar{Y}} = \frac{\sigma}{\sqrt{n}}$

Definition

The standard error of an estimate is the standard deviation of the estimate's sampling distribution.

- Two things govern the standard error ...
  1. How the population varies (σ)
  2. Sample size (n)
- In fact, we seldom know the population standard deviation ($\sigma_{\bar{Y}}$) and so have to work with the sample standard deviation (s) when calculating the standard error

The Standard Error of an Estimate

Definition

The standard error of the mean is estimated from the sample at hand and calculated as ... $SE_{\bar{Y}} = \frac{s}{\sqrt{n}}$

Note: When calculating $SE_{\bar{Y}}$ we divide by n and not by $n-1$

- When $n = 30; s = 1522.082; SE_{\bar{Y}} = \frac{1522.082}{\sqrt{30}} = 277.8929$
- When $n = 60; s = 1522.082; SE_{\bar{Y}} = \frac{1522.082}{\sqrt{60}} = 196.4999$
- When $n = 100; s = 1522.082; SE_{\bar{Y}} = \frac{1522.082}{\sqrt{100}} = 152.2082$
- Of course, if σ is large then so will be s and as a result so will be $SE_{\bar{Y}}$
- Note also that every estimate (Median, correlation coefficient, etc.) has a standard error associated with it

Confidence Intervals

- Since we do not see the population and have a single estimate drawn from the sample (say, $\bar{Y}$), how sure can we be that we are close to μ?
- Confidence Intervals help us answer this question

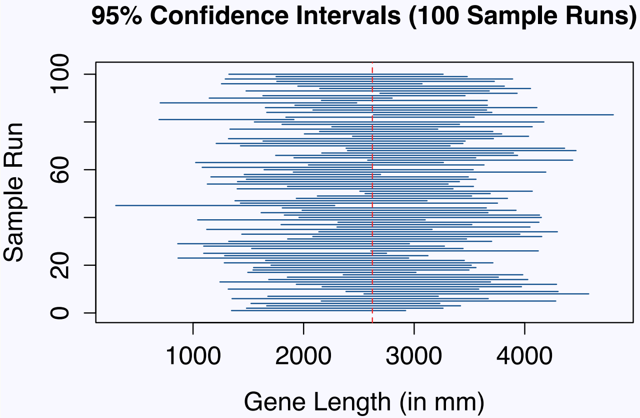
Definition

... A range of plausible values that surround the sample estimate and this range of plausible values is likely to contain the population parameter

- Confidence intervals typically used: 95% or 99%, and you hear folks say "we can be 95% confident that the true parameter (for e.g., the population mean) lies between values x and y " [popular phrasing]
- What they should say is that if "we drew all possible samples of size n and calculated the resulting sample estimates, the range of estimates established by 95% of the 95% confidence intervals calculated for the resulting sample means would trap the population mean"
- *Rule of thumb*: 95% confidence interval is $\approx = \bar{Y} \pm 2SE$

Confidence Interval Simulation

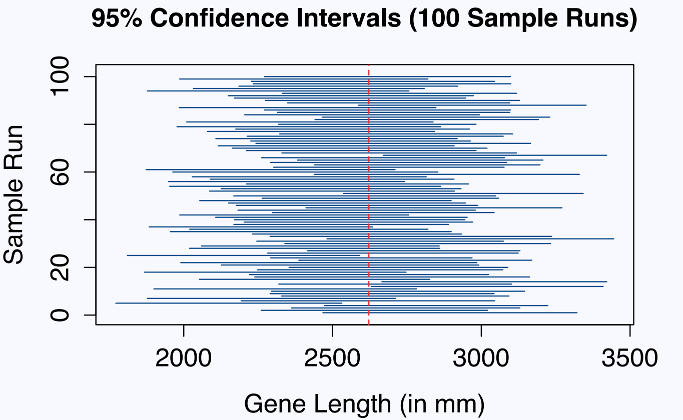
$n = 20$



Note: Only 94 CIs touch $\mu = 2,622$ (the hashed red line)

… once more

$n = 100$



Note: Only 95 CIs touch $\mu = 2,622$ (the hashed red line)

Worked Examples

Worked Example 1

Practice Problem #2

1. The standard error of the mean time to rigor mortis is 0.22 hours (which is approximately 13.27 minutes
2. The standard error measures the spread of the sampling distribution of mean time to rigor mortis
3. That the data represent a random sample of time to rigor mortis

Worked Example 2

Practice Problem #7

1. Mean flash duration is 95.94 milliseconds
2. No, it is very unlikely because this estimates is based upon a small sample of 35 male fireflies
3. The standard error is 1.85 milliseconds
4. The standard error tells us how far, on average, we might expect our sample mean to be from the population mean.
5. The approximate 95% CI is:
   $95.94286 \pm 2(1.858409) = (92.22604, 99.65968)$
6. We can be roughly 95% confident that the true population mean of flash duration lies in this interval of $(92.22, 99.65)$ milliseconds.