STAT 113
Working with Theoretical Distributions

Colin Reimer Dawson

Oberlin College

November 2, 2017

Outline

P-value = Proportion of Randomized Sample Statistics

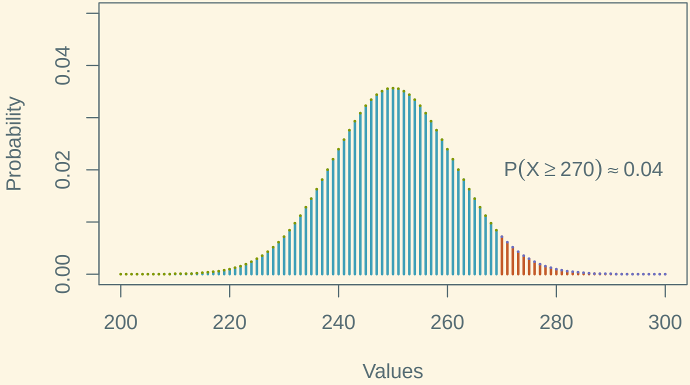


Figure: Randomization distribution for the number of heads in 500 coin flips, highlighting the one-tailed P-value testing $H_1 : p > 0.5$ for an observation of 270 heads.

Confidence Level = Proportion of Bootstrap Samples

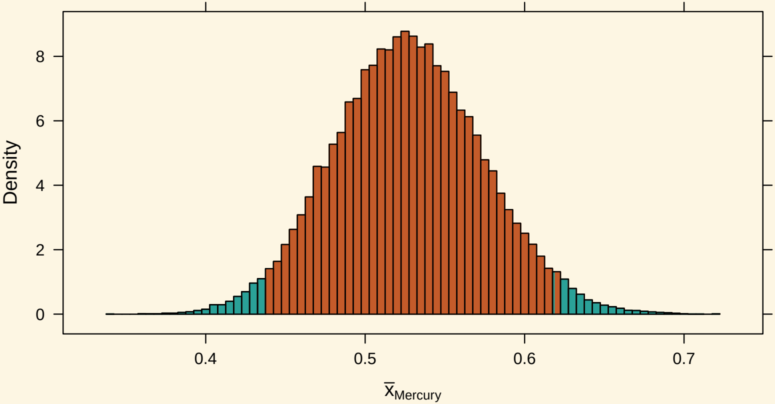


Figure: Bootstrap distribution for mean mercury level in fish in Florida Lakes (from `FloridaLakes` dataset). The middle 95% is highlighted illustrating a 95% confidence interval.

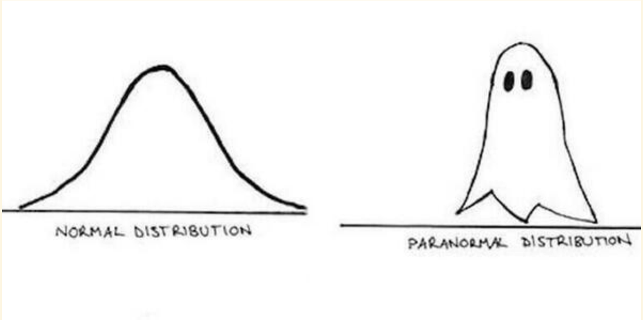
Properties of Sampling Distributions

Most (about 95%) of *simple random* samples have a sample mean ($\bar{x}$) which is within 2 Standard Errors of the population mean (μ). Therefore, about 95% of the time, the population mean will be within 2SE of the *sample* mean!

A similar statement holds for some other statistics/parameters, under a particular condition. What condition? The sampling distribution needs to be (approximately) symmetric and bell-shaped

So what's with all these bell shapes?

- Q: Why are so many distributions "bell-shaped"?
- A: The **Central Limit Theorem**
- One of the most important results in probability: for sufficiently large samples, sample means have a **Normal** (bell-shaped) **distribution**.



Sample Means Show Up A Lot

- Sample means are sample means (did you know this?)
- Sample proportions are sample means (encode binary variable as 0s and 1s)

Even More Stuff is Normal

Also...

- Sum of two Normals is Normal
- Rescaling a Normal by a constant is Normal
- Difference of Normals is Normal

So...

- Sampling distribution for difference of sample means is approximately Normal
- Sampling distribution for difference of sample proportions is approximately Normal

Outline

Analytic Approximations

Density Functions

Properties of Normal Distributions

Normal Bootstrap Distibution

Approximating with a Smooth Curve

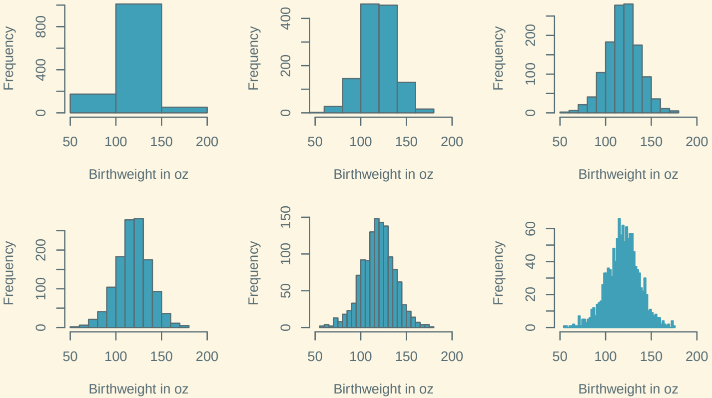


Figure: Frequency Histograms of Babies' Birth Weights (Nolan and Speed, 2000)

Density

$$\text{Proportion} = \text{Area} = \text{Height} \times \text{Width}$$

$$\text{Density} = \text{Height} = \frac{\text{Proportion}}{\text{Width}}$$

This quantity (proportion divided by width) is called “density” by analogy to physics: “amount of stuff” divided by “amount of space”.

Density Histograms

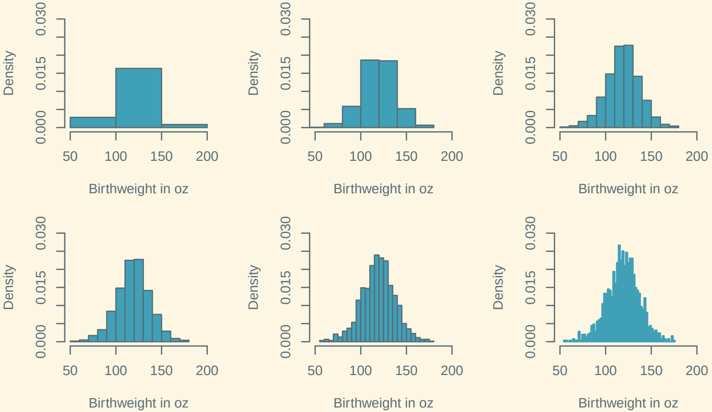


Figure: Density Histograms of Babies' Birth Weights (Nolan and Speed, 2000)

Density Functions

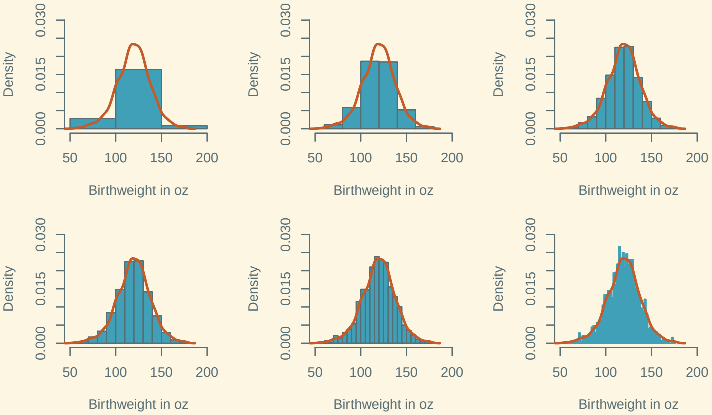


Figure: Densities of Babies' Birth Weights (Nolan and Speed, 2000)

Proportion = Area Under the Density Curve

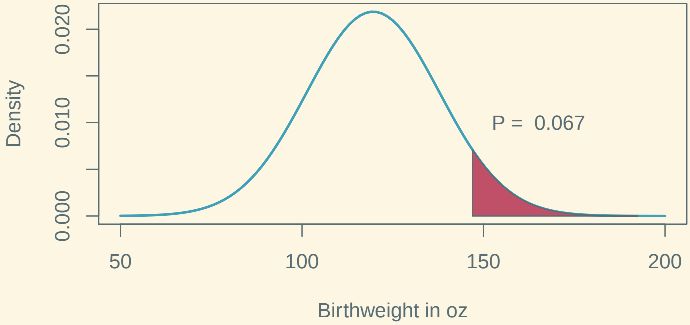
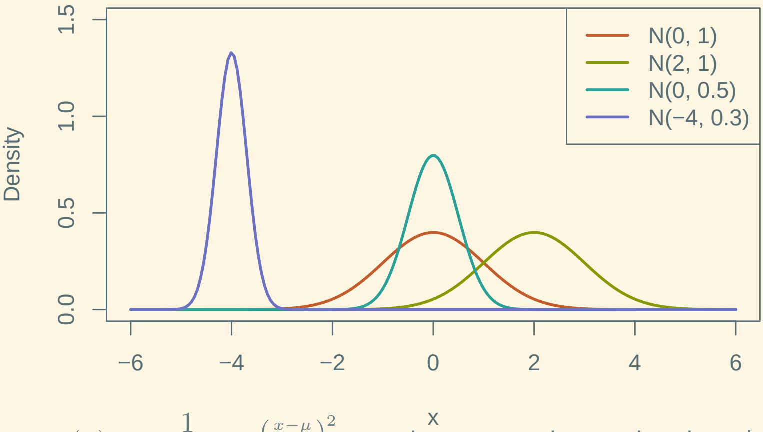


Figure: Approximating birth weight distribution using a Normal. Shaded area is $P(\text{Weight} \geq 148 \text{ oz})$

Outline

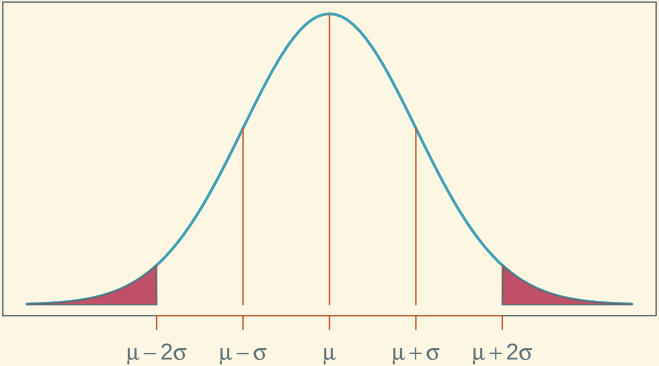
Normal Distributions

Normal distributions are completely specified by their mean (μ) and their standard deviation (σ). We can write $\mathcal{N}(0,1)$ as shorthand for a Normal with mean 0 and standard deviation 1.



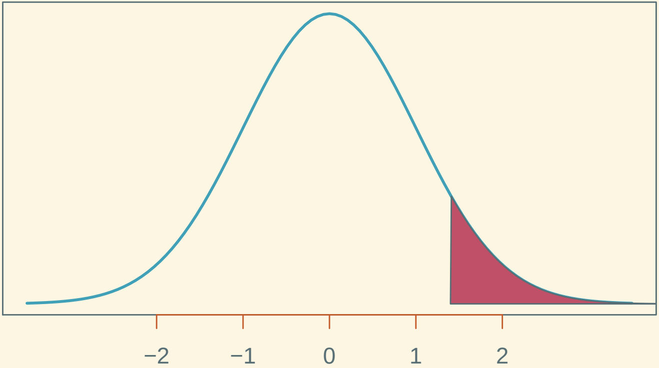
$\text{density}(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-\mu}{\sigma}\right)^2},$ but we won't use this directly.

Normal Distributions



Pairs: (Approximately) what proportion of the area under the curve is shaded? In a bell-shaped (normal) distribution, 95% of cases lie within 2 standard deviations of the mean. So 5% lie beyond 2σ from μ.

Area Under Normal Curve

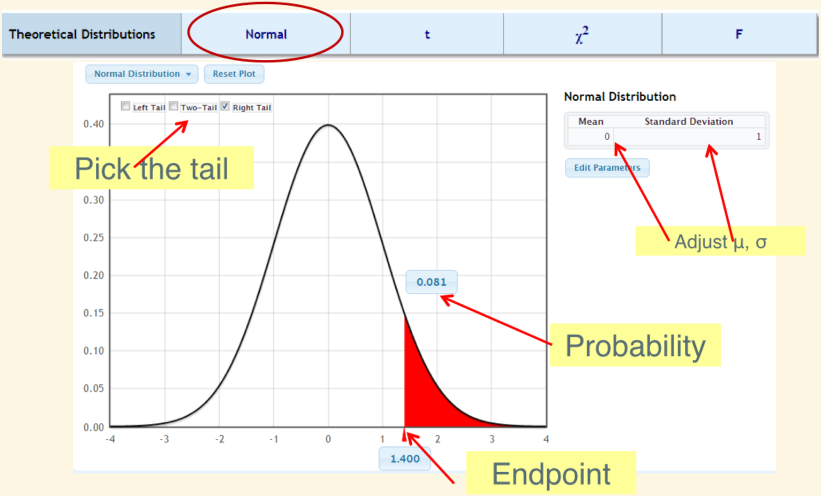


Area under a curve using calculus:

$$\int_{1.5}^{\infty} \frac{1}{\sigma\sqrt{2\pi}} e^{-\left(\frac{x-0}{1}\right)^2} \, dx$$

but this integrand doesn't have a closed-form antiderivative

StatKey to the Rescue!



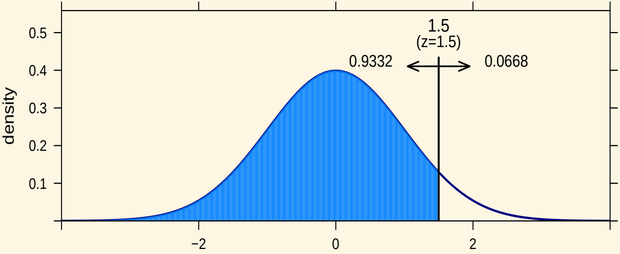
R Works Too

```
library("mosaic")
## Area to the right of 1.5
xpnorm(1.5, mean = 0, sd = 1, lower.tail = FALSE)



    If X ~ N(0, 1), then

     P(X <= 1.5) = P(Z <= 1.5) = 0.9331928
     P(X >  1.5) = P(Z >  1.5) = 0.0668072
```

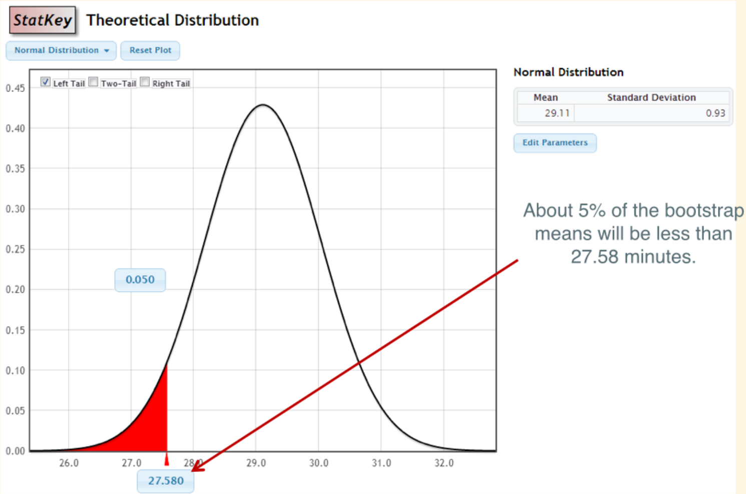


Outline

Quantiles of a Normal Curve

Suppose that the bootstrap distribution of means for samples of size 500 Atlanta commute times is $\mathcal{N}(29.11, 0.93)$. Find an endpoint (percentile) so that just 5% of the bootstrap means are smaller.

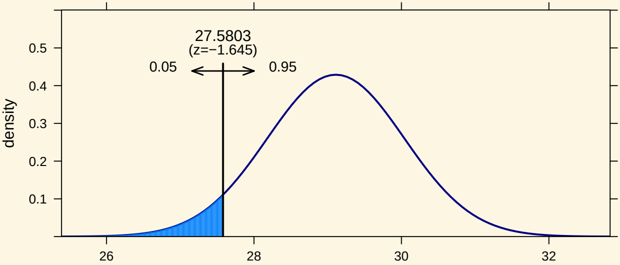
StatKey...


StatKey Theoretical Distribution
Normal Distribution
Reset Plot
Left Tail
Two-Tail
Right Tail
0.45
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.050
26.0
27.0
29.0
30.0
31.0
32.0
27.580
Normal Distribution
Mean
Standard Deviation
29.11
0.93
Edit Parameters
About 5% of the bootstrap means will be less than 27.58 minutes.

And in R ...

```
xqnorm(0.05, mean = 29.11, sd = 0.93)

     P(X <= 27.5802861269351) = 0.05
     P(X >  27.5802861269351) = 0.95
```




```
    [1] 27.58029
```

Goals

Confidence Intervals

If we can approximate a bootstrap distribution with a Normal, we can construct a confidence interval.

P-values

If we can approximate a randomization distribution with a Normal, we can compute P-values.