STAT 215
Inference Concepts Wrap-Up

Colin Reimer Dawson

Oberlin College

September 23rd, 2016

Tests and Confidence Intervals

Testing a Mean

Percentile-Based Intervals

Normal Distributions

Density and Area

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).

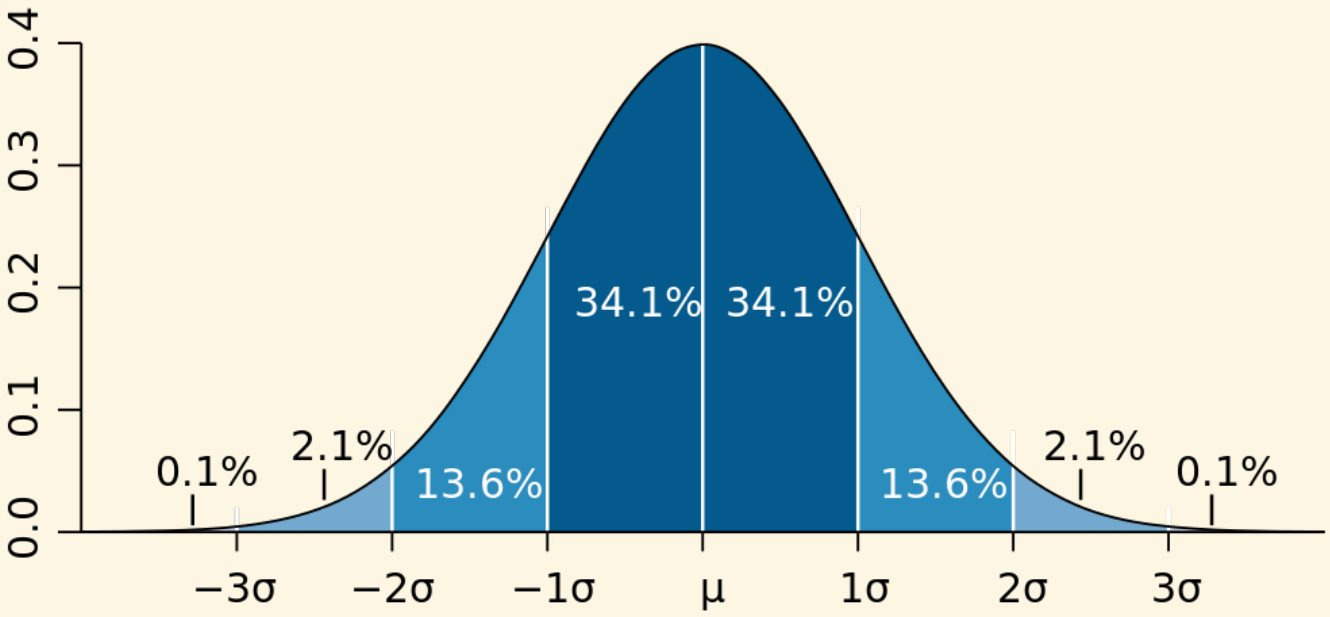


Figure: A Normal Distribution

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

Confidence Level and Significance Level

- We can construct a 95% CI for a population mean, μ, using $\bar{x} \pm 2SE$. We expect that this will capture μ 95% of the time.
- How often will we miss?
- We will miss the other 5% $(1 - C)$ of the time.
- *When* will we miss?
- We will miss when our sample yields a mean which is more than $2SE$ from the population mean.

Confidence Level and Significance Level

- In a two-tailed test for μ with $\alpha = 0.05$, we reject H_0 if the sample mean, $\bar{x}$ falls in the outer 5% of the H_0 sampling distribution of possible $\bar{x}$ values...
- in other words, if $\bar{x}$ is more than $2SE$ from the H_0 value of μ.
- When H_0 is true, how often will we incorrectly reject it?
- We will make a Type I Error 5% (α) of the time that H_0 is true.
- *When* will we incorrectly reject it?
- when the sample statistic happens to be more than 2SE from the population parameter

Confidence Level and Significance Level

- Suppose H_0 is true. How often will the 95% CI contain the H_0 parameter value?
- It will contain the true parameter 95% of the time.
- So if it doesn't...
- Then either H_0 is wrong, or we are in the unfortunate 5% of cases where the distance between μ and $\bar{x}$ is more than $2SE$.
- These are exactly the cases where we *reject* H_0.

Randomization Test for a Single Population Mean

Hypothesis Testing for a Single Population Mean

If the sampling distribution is symmetric and bell-shaped, then we can reject H_0 at $\alpha = 0.05$ if $|\mu_0 - \bar{x}| > 2SE$ (μ_0 is μ according to H_0). We can estimate the SE using a bootstrap distribution.

More generally, estimate the sampling distribution under H_0 by shifting the bootstrap distribution to be centered at μ_0, and compute P-values as for proportion tests.

Alternatively, construct a $(1 - \alpha)$ CI, and reject if μ_0 lies outside it.

Another Way to Interpret CIs

We can think of the $(1 - \alpha)$ Confidence Interval as the set of parameter values that would *not* be rejected using a test with significance level α.

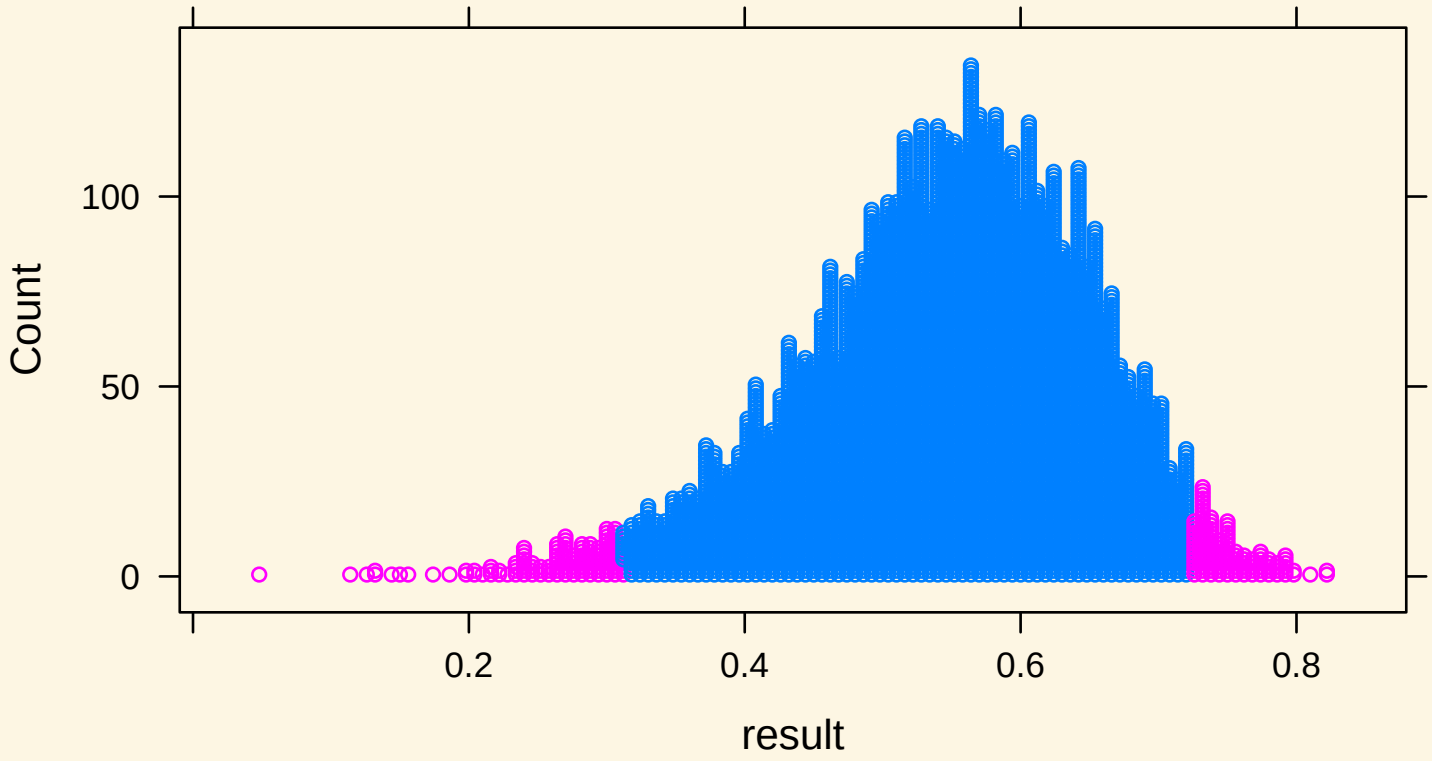
Percentile-Based Intervals

- If the sampling distribution of the statistic is not Normal, it may not be true that 95% of samples are within 2 SE.
- A more general method is to use the bootstrap distribution to estimate the confidence interval directly.

The "Reverse Percentile" Method

```
sample.r <- cor(Price ~ Area, data = Sample60); sample.r
```

```
[1] 0.5462743
```



Estimating the Left and Right MoE directly

```
bounds <- quantile(~result, data = Bootstrap.cors, probs = c(0.025, 0.975))
center <- sample.r
c(boot = bounds[1], sample.r = center, boot = bounds[2]);
```

```
      boot.2.5%   sample.r boot.97.5%
      0.3101024  0.5462743  0.7228145
```

```
c(left = center - bounds[1], right = bounds[2] - center)
```

```
       left.2.5% right.97.5%
       0.2361719   0.1765402
```

We estimate from the bootstrap distribution that 2.5% of the time the population correlation is more than 0.236 *above* the sample correlation about 2.5% of time, and is more than 0.177 *below* the sample correlation about 2.5% of the time.

So, if we construct an interval with a right MoE of 0.236 and a left MoE of 0.177, we should miss the population correlation about 5% of the time.

Constructing the Reverse Percentile Bootstrap Interval

```
RP.lower <- sample.r - (bounds[2] - center)
RP.upper <- sample.r + (center - bounds[1])
c(Lower = RP.lower, Upper = RP.upper)
```

```
     Lower.97.5%  Upper.2.5%
       0.3697340   0.7824461
```

Compare:

```
se.hat <- sd(~result, data = Bootstrap.cors)
Norm.lower <- sample.r - 1.96 * se.hat; Norm.upper <- sample.r + 1.96 * se.hat
c(Lower = Norm.lower, Upper = Norm.upper)
```

```
     Lower     Upper
0.3407813 0.7517672
```

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)

Also Sums, Differences, Rescalings, ...

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

So...

- Differences of sample means are approximately Normal
- Differences of sample proportions are approximately Normal

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

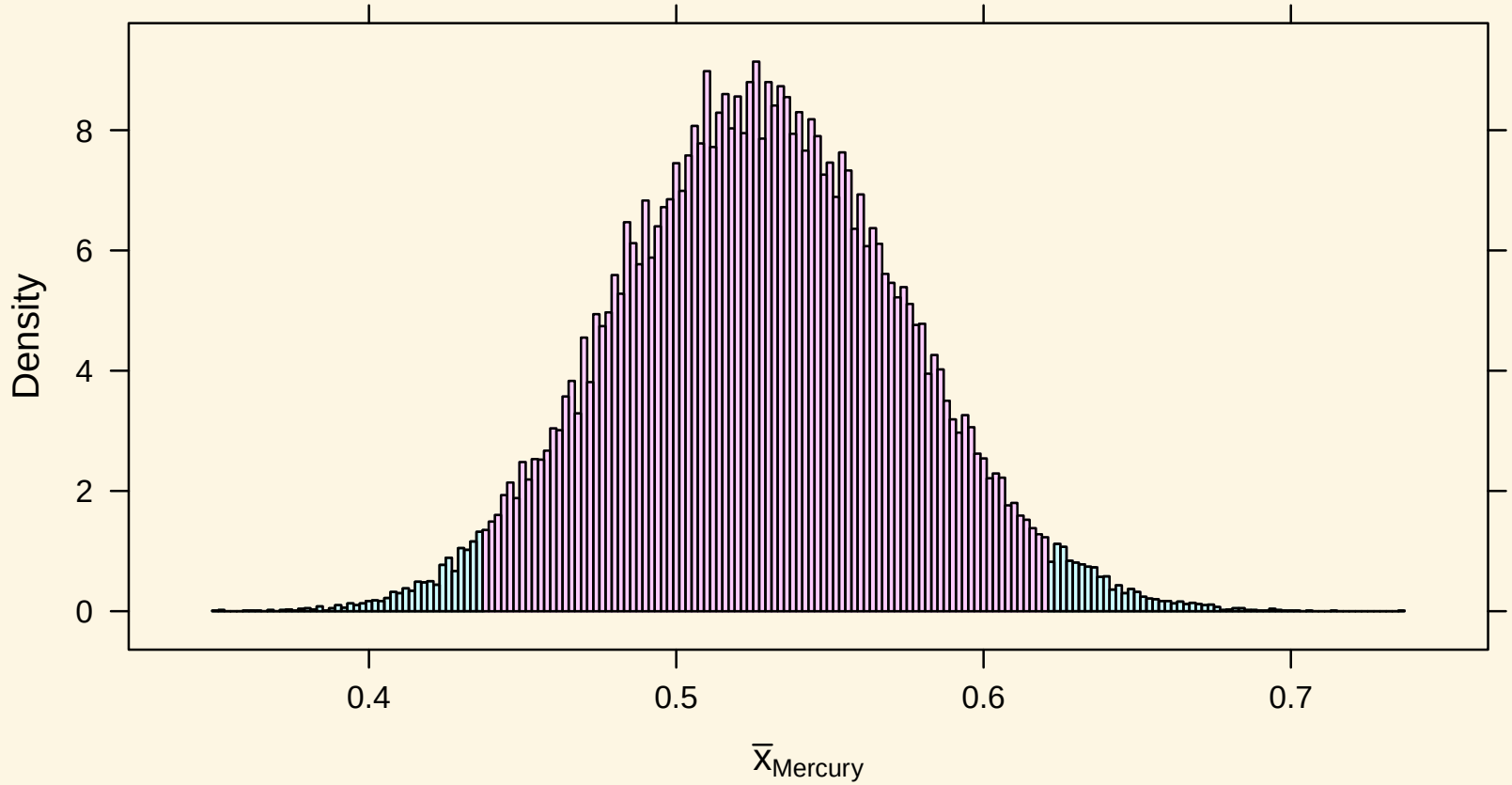


Figure: Bootstrap distribution for mean mercury level in fish in Florida Lakes (from `FloridaLakes` dataset). The middle 95% is highlighted illustrating a 95% confidence interval.

Narrowing Bins: Frequency

Figure: 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".

Narrowing Bins: Density

Figure: 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 Sampling Distribution of $\bar{X}$ using a Normal. Shaded area is $P(\bar{X} > 103)$

Analytic Approximations of Sampling Distributions

Param.	Stat.	Randomization	Theory SE	Test Dist.
p	$\hat{p}$	Simulate from p_0	$\sqrt{\frac{p_0(1-p_0)}{n}}$	Normal
μ	$\bar{x}$	Bootstrap + shift	$\frac{s}{\sqrt{n}}$	t_{n-1}
$p_A - p_B$	$\hat{p}_A - \hat{p}_B$	Scramble groups	$\sqrt{\frac{\hat{p}(1-\hat{p})}{n_A} + \frac{\hat{p}(1-\hat{p})}{n_B}}$	Normal
$\mu_A - \mu_B$	$\bar{x}_A - \bar{x}_B$	Scramble groups	$\sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}$	$t_{\min(n_A-1, n_B-1)}$
μ_D	$\bar{x}_D$	Flip pairs*	$\frac{s_D}{\sqrt{n_D}}$	t_{n_D-1}
ρ	r	Scramble pairings	$\sqrt{\frac{1-r^2}{n-2}}$	t_{n-2}