Day 2, Morning:
The Logic of Distributions

Instructor: Marshall A. Taylor
844 Flanner Hall
mtaylo15@nd.edu
marshalltaylor.net

Recap

- Yesterday morning we talked about:
  - The descriptive and inferential purposes of statistics
  - The difference between samples, populations, and the issues that arise because of sampling error and sample bias—and the ways in which probability theory can be used to address the former.
  - The basics of probability theory

Recap

- Yesterday evening we talked about:
  - The difference between univariate, bivariate, and multivariate statistics.
  - What a variable is and the general forms that it can take: nominal, ordinal, or continuous.
  - Measures of central tendency for each type of variable.
  - Measures of dispersion.

Game Plan for Today

- Morning
  - We will bring these previous lectures together to show how we can use probability theory to assess how representative our **sample distribution** is of the **population distribution**.

- Evening
  - After outlining the **asymptotic theory of probability distributions** in the morning, we will then examine basic **univariate tests of statistical inference** to quantify how well our sample statistics approximate population parameters net of sampling error.

What is a distribution?

In statistics, a **distribution** is simply the array of values for one or more variables across a set of units (people, groups, etc.).

The distribution is everything

- The concept of “distribution” has been at the implicit center of absolutely everything we have talked about so far!

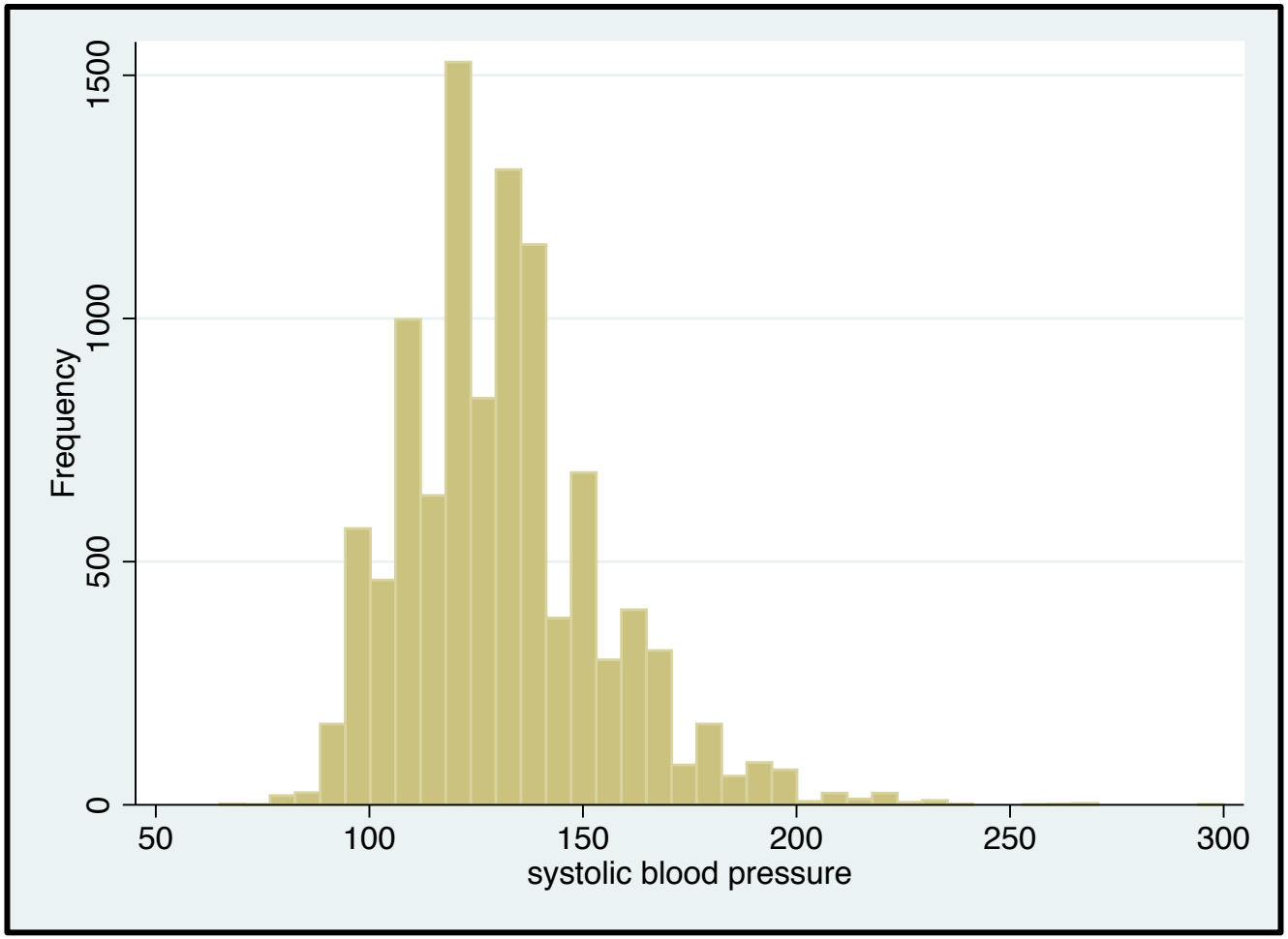
- We have specifically looked at sample statistic distributions, such as…

Frequency Distributions

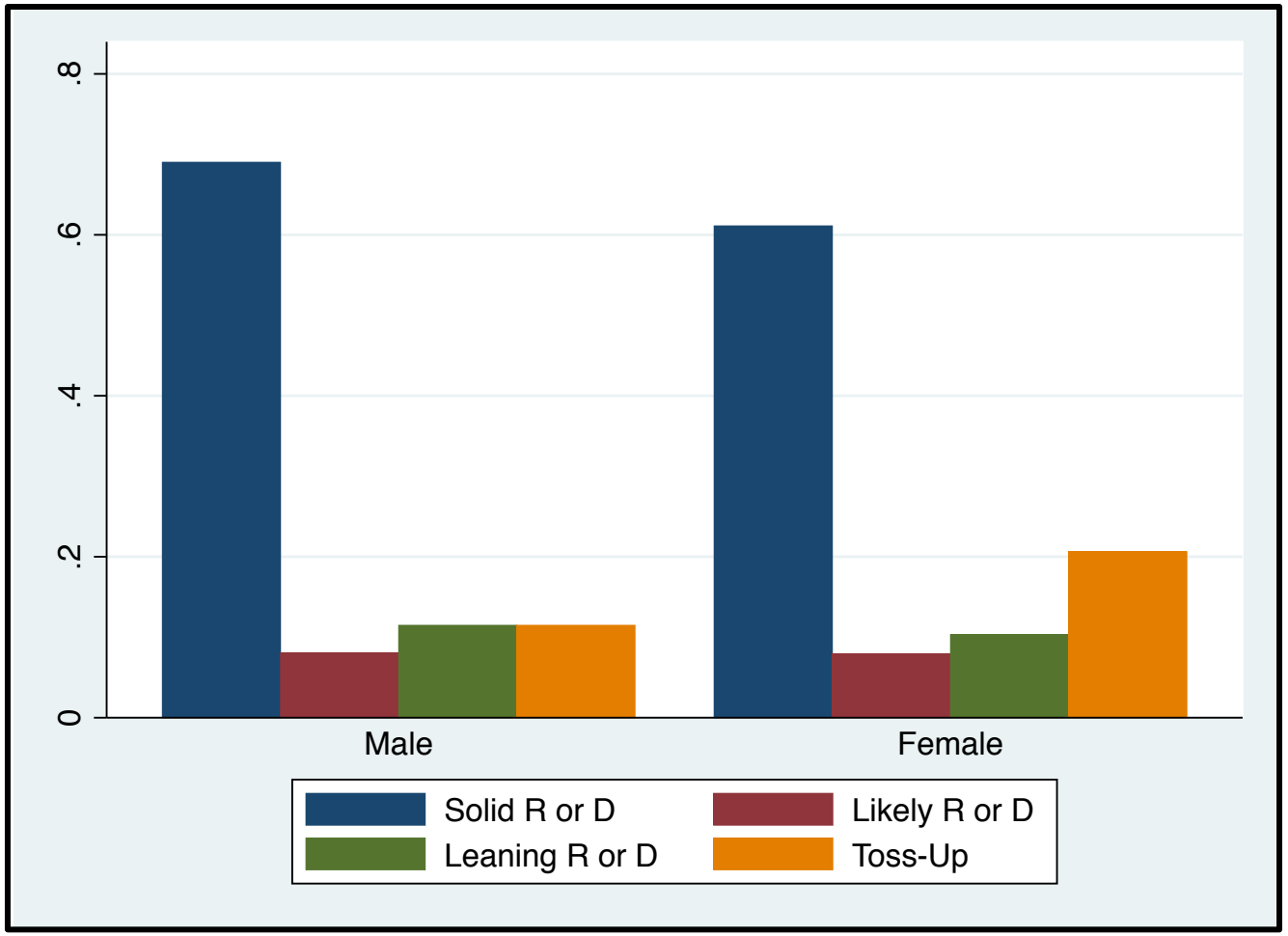
```
. tab health

1=poor,...,
5=excellent |      Freq.     Percent        Cum.
------------+-----------------------------------
       poor |        729        7.05        7.05
       fair |      1,670       16.16       23.21
    average |      2,938       28.43       51.64
       good |      2,591       25.07       76.71
  excellent |      2,407       23.29      100.00
------------+-----------------------------------
      Total |     10,335      100.00
```

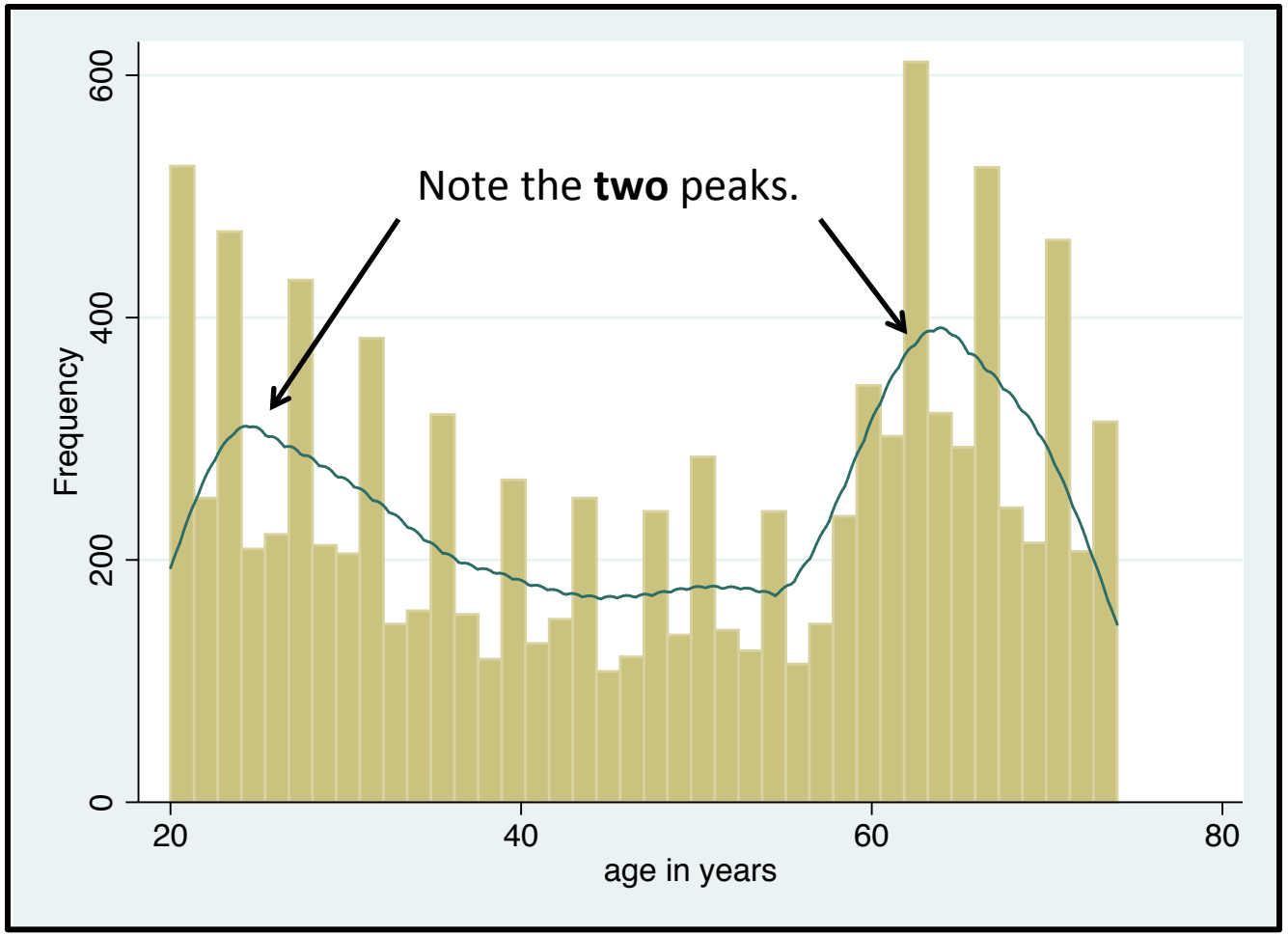
Frequency distributions as "histograms"



Bimodal distribution with categorical variables



Bimodal distribution with continuous variables



Distributions and Probability Theory

- Distributions serve much more than a descriptive purpose.

- We rely on the **asymptotic theory of probability distributions** to make **statistical inferences**.

- Such a theory gives us dependable ideas about what the **sampling distribution** will look like.

What is “asymptotic”?

- “Asymptotic” refers to the property that, if sampled an infinite number of times, a statistic will converge to the population parameter it is meant to approximate.

- That is, $\bar{Y}_n \rightarrow \mu_N$ as $n \rightarrow \infty$, assuming that all n are random samples of N.

What is a “probability distribution”?

- A “probability distribution” is an array of probabilistic values for a variable across a set of units, where the values are proportions that must sum to 1.

	V1	V2	V3	V4	V5	R Margins
1	0.0000576938	0.1088889841	0.1090984138	0.0000576950	0.7818972133	1
2	0.0000476265	0.0000476352	0.9123043912	0.0119956500	0.0756046971	1
3	0.0000881123	0.0179937158	0.0264640660	0.0000881160	0.9553659899	1
4	0.1108000716	0.2503997467	0.5927814401	0.0459855150	0.0000332266	1
5	0.2411895005	0.0404012131	0.6524702741	0.0001041705	0.0658348418	1
6	0.1470047772	0.0002170118	0.6761725881	0.0002169660	0.1763886570	1
7	0.8856909695	0.0001638012	0.0001637739	0.0001637537	0.1138177018	1
8	0.2807926994	0.2093476438	0.0530959614	0.0000664792	0.4566972163	1
9	0.5049335568	0.0260158232	0.0000863853	0.4424611335	0.0265031011	1
10	0.0203117985	0.3800703244	0.2362073342	0.0352103566	0.3282001863	1
11	0.1901308947	0.5532532962	0.0002811341	0.0002811777	0.2560534974	1
12	0.3384214744	0.4279357679	0.1207937032	0.1128139993	0.0000350552	1
13	0.8520671572	0.0001154451	0.0462275639	0.1014744017	0.0001154321	1
14	0.8163828385	0.1158173532	0.0000841640	0.0213096942	0.0464059501	1
15	0.4869992267	0.2536825278	0.0000751098	0.0210530230	0.2381901127	1

What is a "probability distribution"?

$$\sum_{v=1|v_1\cdots,v_n\in r_1}^{V} p_v = 1$$

	v1	v2	v3	v4	v5	R Margins
r1	0.0000576938	0.1088889841	0.1090984138	0.0000576950	0.7818972133	1
r2	0.0000476265	0.0000476352	0.9123043912	0.0119956500	0.0756046971	1
r3	0.0000881123	0.0179937158	0.0264640660	0.0000881160	0.9553659899	1
r4	0.1108000716	0.2503997467	0.5927814401	0.0459855150	0.0000332266	1
r5	0.2411895005	0.0404012131	0.6524702741	0.0001041705	0.0658348418	1
r6	0.1470047772	0.0002170118	0.6761725881	0.0002169660	0.1763886570	1
r7	0.8856909695	0.0001638012	0.0001637739	0.0001637537	0.1138177018	1
r8	0.2807926994	0.2093476438	0.0530959614	0.0000664792	0.4566972163	1
r9	0.5049335568	0.0260158232	0.0000863853	0.4424611335	0.0265031011	1
r10	0.0203117985	0.3800703244	0.2362073342	0.0352103566	0.3282001863	1
r11	0.1901308947	0.5532532962	0.0002811341	0.0002811777	0.2560534974	1
r12	0.3384214744	0.4279357679	0.1207937032	0.1128139993	0.0000350552	1
r13	0.8520671572	0.0001154451	0.0462275639	0.1014744017	0.0001154321	1
r14	0.8163828385	0.1158173532	0.0000841640	0.0213096942	0.0464059501	1
r15	0.4869992267	0.2536825278	0.0000751098	0.0210530230	0.2381901127	1

Putting it together

- If we have an infinite number of random sample estimates from the same population, the mean of this distribution of estimates will converge to the population mean.

- Given that we can never really have an infinite number of samples, the asymptotic theory of probability distributions suggests that, with larger sample sizes, we can **infer** with greater degrees of probabilistic confidence whether or not our single sample statistic accurately reflects the unknown population parameter.

- As n approaches N, the sampling error gets smaller and smaller, meaning that the reliability of our estimate gets better and better. This is because, theoretically, if the sample size (n) keeps growing, it will eventually just be the population (N)!

Putting it together

- Of course, such a theory requires that we make assumptions about the shape of the unknown population distribution.

- Otherwise we don't know what n is approximating!

Central Limit Theorem

- Lucky for us, some very intelligent statisticians who came before us noticed that, as sample sizes grew larger and larger, the distribution of sample means becomes approximately **normal**—regardless of whether or not the parameter itself is normally distributed. **This is the Central Limit Theorem (CLT)**.

- By **normal distribution**, we mean a **symmetric** distribution where approximately half of the data fall to either side of the mean. It is commonly known as a distribution that follows a **bell curve**.

Central Limit Theorem

According to the CLT, we can expect that, with a normal distribution of random sample means, approximately **68%** of the sample means will be within **one** standard deviation on either side of the population mean (μ). **95%** will be within **two**, and **99.7%** within **three**.

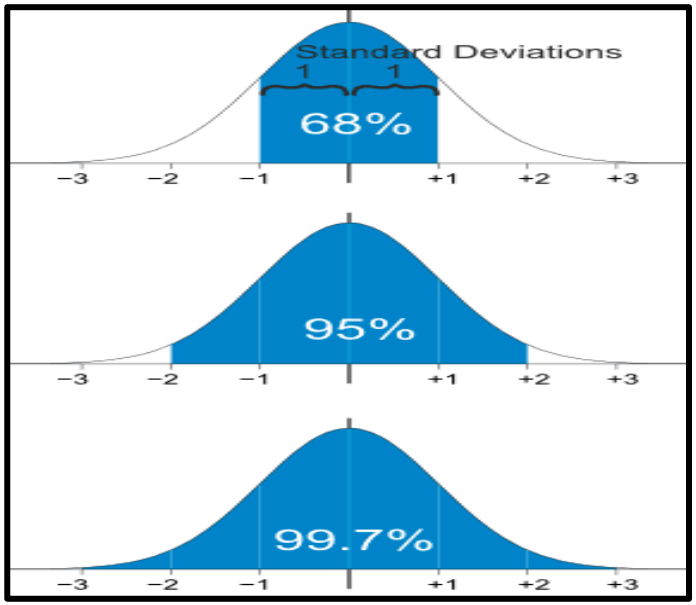
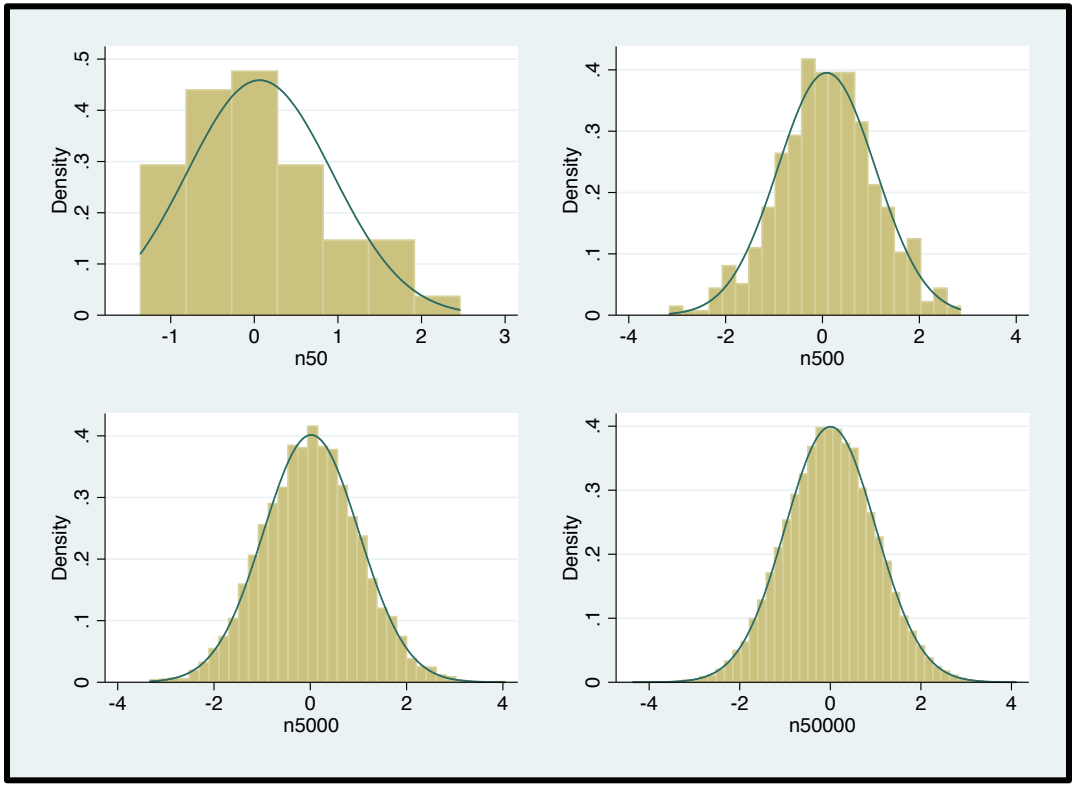


*Figure from MathIsFun website (https://www.mathsisfun.com/data/standard-normal-distribution.html).

Central Limit Theorem

Notice how the data become more symmetric about the mean as the sample size increases. As such, large sample sizes can serve as “proxies” for repeated random samples and justify the CLT.



Standard Error

- So, what can we say about population parameters given a sample statistic and these population distribution assumptions?
- For starters, we can calculate the standard deviation of the theoretical distribution of random sample means around the unknown population mean—also known as the standard deviation of the sampling distribution. This is known as the **standard error**, and can be found with:

$$\sigma(\bar{Y}) = \frac{\sigma}{\sqrt{n}}$$

  - Where σ is the standard deviation of the population parameter and the denominator is the square root of the sample size.

Standard Error

- Of course, σ is usually not known, so we use the sample standard deviation as an approximation:

$$\sigma(\bar{Y}) = \frac{1}{\sqrt{n}}\sqrt{\frac{\sum_{i=1}^{n}(Y - \bar{Y})^2}{n-1}}$$

- Or more simply:

$$\sigma(\bar{Y}) = \frac{s}{\sqrt{n}}$$

Standard Error

- What does the standard error of the systolic blood pressure variable tell us? How is this difference from the standard deviation?

```
. sum bpsystol

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    bpsystol |      10337    130.8826    23.34159        65        300

. ci bpsystol

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    bpsystol |      10337    130.8826    .2295796        130.4325    131.3326
```

Standard Error

- A random sample mean drawn from the population (such as this one) likely differs from the population systolic bp by about 0.23 mm/Hg.

```
. sum bpsystol

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    bpsystol |      10337    130.8826    23.34159        65        300

. ci bpsystol

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    bpsystol |      10337    130.8826    .2295796        130.4325    131.3326
```

Standard Error

- The standard deviation, however, is merely a descriptive indication of variable dispersion. The average respondent in the sample diverges about 23.34 mm/Hg. from the mean.

```
. sum bpsystol

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
    bpsystol |     10337    130.8826    23.34159        65        300

. ci bpsystol

    Variable |       Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    bpsystol |     10337    130.8826    .2295796        130.4325    131.3326
```

Standard Error

- Just to check the math:

$$\sigma(\bar{Y}) = \frac{23.34159}{\sqrt{10,337}} = .2295796$$

```
. sum bpsystol

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
    bpsystol |      10337    130.8826    23.34159        65        300

. ci bpsystol

    Variable |        Obs        Mean    Std. Err.       [95% Conf. Interval]
-------------+---------------------------------------------------------------
    bpsystol |      10337    130.8826    .2295796        130.4325    131.3326
```

Standard Error

- Notice that $\sigma(\bar{Y})$ gets **smaller** when two things happen:
  - (1) When s, the standard deviation, is **small**.
  - (2) When the sample size is **large**.

- But also note that s itself is smaller when the sample size is larger.

- **When n is large—and therefore a closer approximation of N—the sampling distribution varies less and it is more likely that the sample mean represents the population mean!**

Confidence Interval for Mean

- We can also use the standard error and our knowledge of the normal distribution to construct a **confidence interval around the mean**—that is, the band of values within which the population mean, μ, is likely to reside.

Confidence Interval for Mean

- The confidence interval can be found with:

$$\mathrm{CI} = \bar{Y} \pm z(\frac{s}{\sqrt{n}}) \quad \text{or} \quad \mathrm{CI} = \bar{Y} \pm z * \sigma(\bar{Y})$$

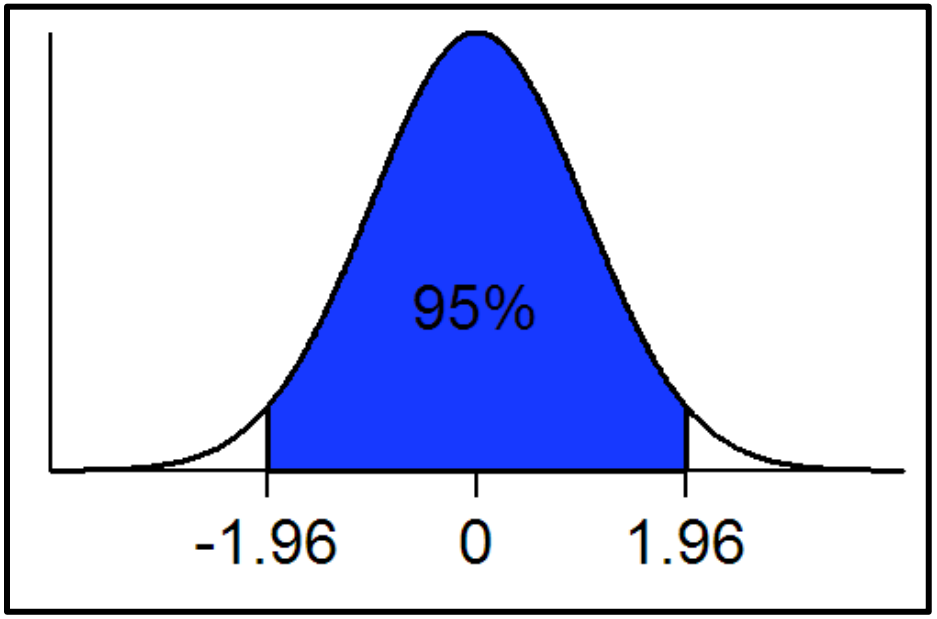
- Where z is our **critical value**: i.e., the number of standard deviations away from the mean that represent the range of probabilities within which we think the population mean resides.

Wait... z-value? What's that?

- Think back to what the CLT tells us:
  - About **68%** of sample means fall within *about* **one** standard deviation on either side of the population mean.
  - About **95%** fall within *about* **two**.
  - About **99.7%** fall within *about* **three**.
- We can use this information to find the standard deviations that correspond to the distribution percentiles that capture these percentages.

Wait… z value? What's that?

- For example, though we say that 95% of the estimates fall within about two standard deviations of the population mean, the more precise number is **1.96**. It is our z **value**!



That is, about **95%** of the sample means fall within **± 1.96 standard deviations** of the population mean. (Never mind that 0—for our purposes, think of it as μ.)

*Photo from Wikipedia (https://en.wikipedia.org/wiki/1.96).

Confidence Interval Example

- The mean weight (in kilograms) in our NHANES sample is 71.90. The standard deviation is 15.36, and our sample size is 10,337. Within what range of kilograms can we be 95% confident includes the population mean?

```
. sum weight

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      weight |     10337    71.90088    15.35515       30.84     175.88
```

Confidence Interval Example

- Let's start by first computing the standard error:

$$\sigma(\bar{Y}) = \frac{s}{\sqrt{n}} = \frac{15.35515}{\sqrt{10,337}} = .15102777$$

```
. sum weight

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      weight |     10337    71.90088    15.35515       30.84     175.88
```

Confidence Interval Example

- We want to capture the population mean within the band of values that, according to the CLT, likely fall within ±1.96 standard deviations from the population mean. As such:

$$\text{CI} = \bar{Y} \pm z(\frac{s}{\sqrt{n}}) = 71.901 \pm (1.96)(.151) = 71.604 \leq \mu \leq 72.197$$

```
. sum weight

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      weight |      10337    71.90088    15.35515       30.84     175.88
```

Confidence Interval Example

- There is a 95% chance that the interval between 71.604 kg. and 72.197 kg. contains the mean population weight.

$$\text{CI} = \bar{Y} \pm z(\frac{s}{\sqrt{n}}) = 71.901 \pm (1.96)(.151) = 71.604 \leq \mu \leq 72.197$$

```
. sum weight

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      weight |     10337    71.90088    15.35515       30.84     175.88
```

Confidence Interval Example

- Confirm with Stata:

Variable	Obs	Mean	Std. Err.	[95% Conf.	Interval]
weight	10337	71.90088	.1510277	71.60484	72.19692

CI example with different critical value

- What if we wanted to be, say, 99% confident that our interval contains μ?

- The critical z-value for a 99% confidence interval is **2.58**. This means that, following the CLT, we expect about **99%** of sample means pulled randomly from our sampling distribution fall within ± 2.58 standard deviations of the population mean.

CI example with different critical value

- Let's do the math:

$$\text{CI} = \bar{Y} \pm z(\frac{s}{\sqrt{n}}) = 71.901 \pm (2.58)(.151) = 71.511 \leq \mu \leq 72.291$$

```
. sum weight

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
      weight |      10337    71.90088    15.35515       30.84     175.88
```

CI example with different critical value

- We can say that, 99 times out of 100, we have captured the mean population weight with the interval between 71.51 kg. and 72.29 kg.

$$\text{CI} = \bar{Y} \pm z(\frac{s}{\sqrt{n}}) = 71.901 \pm (2.58)(.151) = 71.511 \leq \mu \leq 72.291$$

```
. sum weight

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
      weight |     10337    71.90088    15.35515       30.84     175.88
```

CI example with different critical value

- Confirm with Stata:

Variable	Obs	Mean	Std. Err.	[99% Conf.	Interval]
weight	10337	71.90088	.1510277	71.51179	72.28997

Confidence Interval Precision

- Note that the confidence interval gets **bigger** when we go from 95% to 99% confidence.
  - For 95% CI: 72.197 – 71.605 = **.592**
  - For 99% CI: 72.290 – 71.512 = **.778**

- This is because we have to have **less precision** when we try to be more confident!

z-scores

- Recall that critical z-values (e.g., ±1.96 and ±2.58) are the standard deviations away from μ that we would expect to capture 95% and 99% of random sample means (respectively) in a normal sampling distribution.

- Theoretically, the z-value for any given case can be calculated with $(Y - \mu)/\sigma$. This value would tell us how many standard deviations the case is from μ.

- We can apply this same logic to an individual variable to quantify how far a specific case is from the **variable mean**. This measure is called a **z-score**, or a **standardized score**.

z-scores

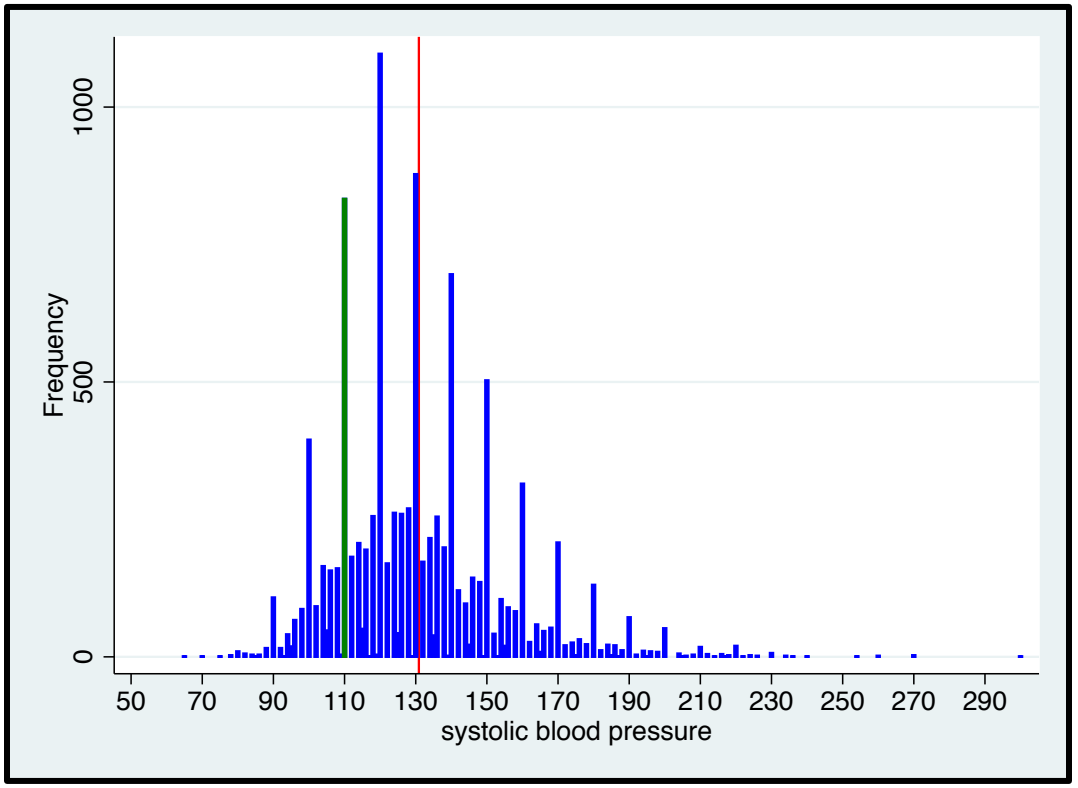
- The z-score for an individual case can be found by subtracting the variable mean from the raw score and then dividing the difference by the variable standard deviation:

$$\frac{(Y - \bar{Y})}{s}$$

  - Where, as before, $\bar{Y}$ is the mean for the variable and s is the variable standard deviation.

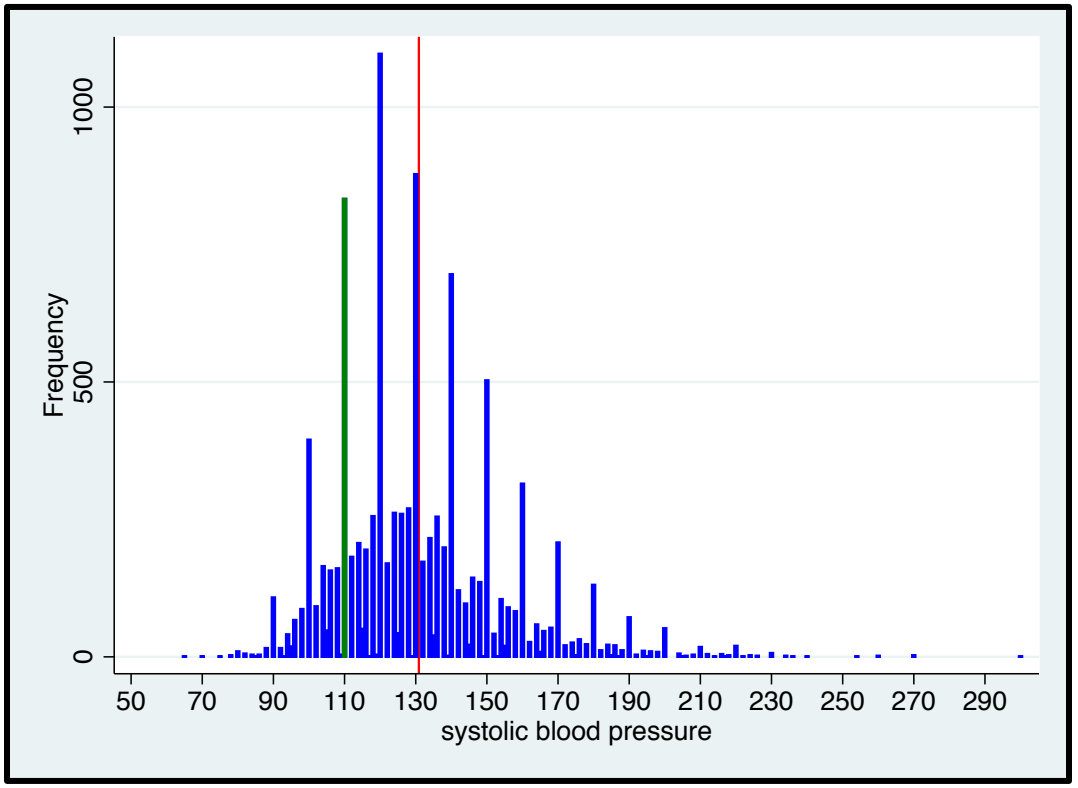
z-score example

- Below is the distribution of systolic blood pressure readings for the NHANES sample. The mean is 130.88 mm/Hg. The green bar is a particular value of the variable: 110 mm/Hg. The standard deviation is 23.34 mm/Hg. What is the z-score for this case, and what does this number mean?



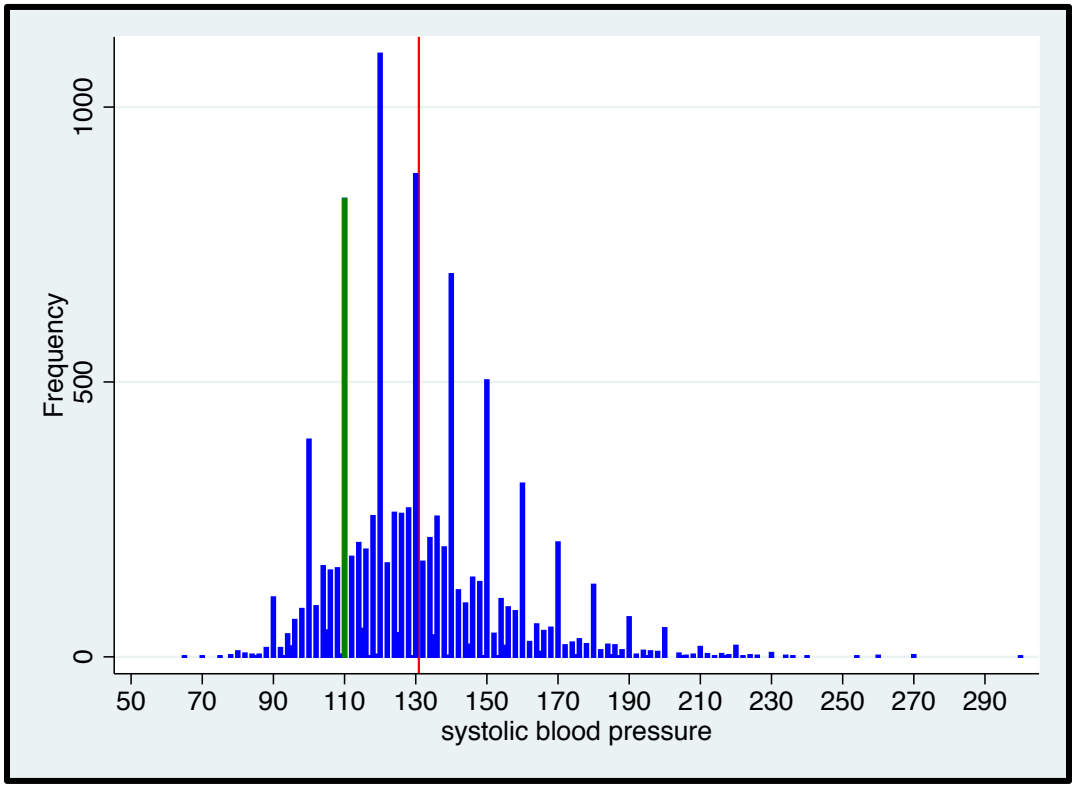
z-score example

$$z = \frac{(Y - \bar{Y})}{s} = \frac{110 - 130.8826}{23.34} = -0.895$$



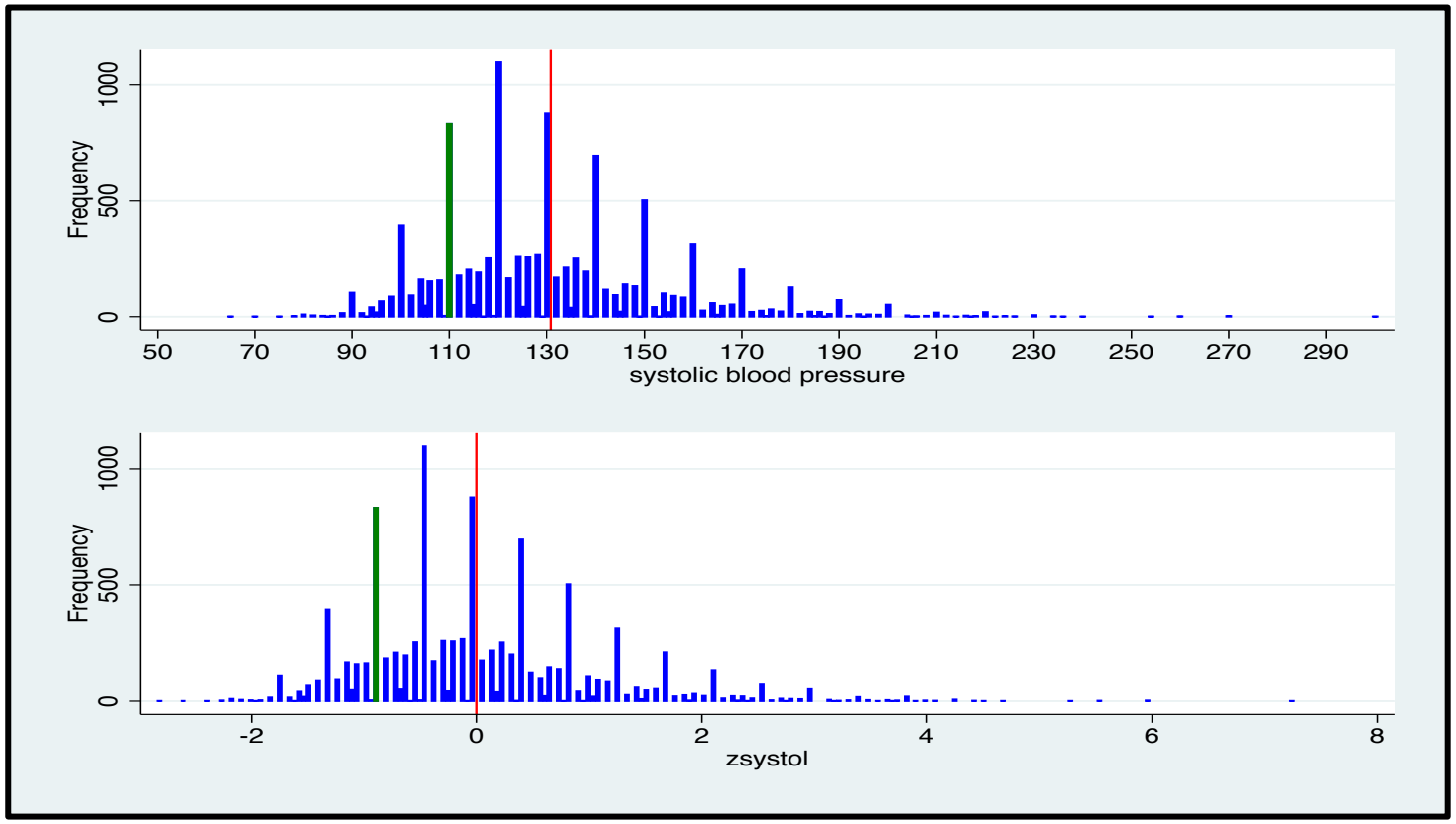
z-score example

- A case with a systolic blood pressure reading of 110 mm/Hg. is a little less than 1 standard deviation **below** the mean.



z-score example

- Confirming with Stata. Note that the same bar is colored green. That's because they are the same cases!



Conclusion

- We have seen how the asymptotic theory of probability distributions allows us to assess how well our single sample mean represents the true population mean in the absence of repeated random samples.

- It does this by following the CLT. This allows us to use our sample size to estimate how well hypothetical random samples (of the same size) would approximate a normal distribution and therefore approximate the population mean.

Conclusion

- Though standard errors and confidence intervals help us get an idea of where the population mean may be, how do we know these numbers are reliable? That is, how do we know that our estimates aren't just the product of sampling error?

- This is the job of **statistical inference**—and it is the topic for the next session!

Datasets Used

- The Stata survey documentation data, `nhanes2f`, from the *Stata Press* website. Retrieved July 24, 2016 (http://www.stata-press.com/data/r11/svy.html).