Biost 517 / Biost 514 Applied Biostatistics I / Biostatistics I

Timothy A. Thornton, Ph.D.

Associate Professor of Biostatistics

University of Washington

Lecture 5:

Introduction to Statistical Inference; Point Estimates; Confidence Intervals

Description and Inference

- So far we have discussed descriptive statistics: summaries of the data that describe the data.
- We now focus on inferential statistics
  - Inferential statistics can support claims about the population or the underlying process
- Descriptive statistics require choices. Inferential statistics require assumptions.
  - The assumptions can be as minimal as the assumption that the data are a random sample
  - Assumptions can be stronger: If we use a Kaplan-Meier curve to estimate the survival curve for a population, noninformative censoring is assumed

Assumptions

- It is important to distinguish *necessary* and *sufficient* assumptions.
  - Noninformative censoring is *necessary* in order for Kaplan-Meier survival curves to estimate the survival curve in the population
  - Some statistics textbooks say that to make inference about a population mean, we assume the distribution of the data in the population has a Normal distribution. This assumption is *sufficient*, but not necessary because of the Central Limit Theorem. An alternative assumption would be that the sample sizes are sufficiently large such that the sample mean is normally distributed, will do just as well.

Statistical Inference

- The purpose of **statistical inference** is to draw conclusions about a population without the full population's data.
- Typically there will be data on a small subset of individuals selected from the population, i.e., a **sample**
- With statistical inference, we are often interested in using statistical inference for substantiating or validating conclusions about the population from the the sample

Parameters

- Summary measures on a population are called *parameters*
- Population parameters: mean height, difference in mean height between men and women, median IQ, the proportion of practicing Catholics, the standard deviation of blood pressure, the correlation between weight and height in a population, the efficacy of the flu vaccine in preventing flu
- We often use summary measures calculated on samples to *estimate* population parameters
- Any inference about a population parameter from the sample must take into account the uncertainty and randomness of the sample.

Sampling Distribution

- The **sampling distribution of a statistic** is the distribution of the values of the statistic in all possible samples of the same size *n* taken from the same population.
- If we were to repeat the study a large number of times (under the exact same conditions) …
  - What would be the distribution of the statistic computed from the samples obtained?
- We need to know the sampling distribution of a statistic to determine if it is appropriate to use for inference about a population parameter

Bias and Variability of Statistics

- Properties of expectation and variance allow us to describe the general tendencies of statistics
  - Unbiasedness: If we repeated the study many times, would the resulting statistics tend to be centered on the true population parameter?
    - More exactly: if our statistic $T=T(X_1,\ldots X_n)$ estimates a parameter θ, then the statistic is unbiased if $E[T]= \theta$
  - Variability: How variable would the estimates be across replicated studies?

Standard Errors of Statistics

- Descriptively, we find standard deviations preferable to variances
  - Variance is in squared units
  - Standard deviation is in the original units
- Because the units are the same, standard deviation is the right choice to describe properties of distributions
  - E.g., 95% of a Normal distribution is within 1.96 SDs of the mean
- "Standard error" is the square root of the variance of a statistic
  - The major motivation for the nomenclature is to distinguish it from the standard deviation in the population of measurements

Properties of Standard Errors

- Derived from variances:
- We compute standard errors by first finding the sampling variance of the statistic
  - "sampling variance" means "variance of the sampling distribution of the statistic"
- When transforming or combining statistics
  - Convert standard errors to variances by squaring
  - Use properties of variances
  - Convert resulting variance back to a standard error

Types of inference

- Point estimation: what is the most likely value for this parameter in the population?
- Interval estimation: a range of values for a parameter that is likely to contain the true population value
- (later) Hypothesis testing: Are the data consistent with the true population value being the value we hypothesized?

Inference for a population mean

- It is often of interest to draw inference about a population mean for a variable
- From lecture 4, we showed that for a random sample of size *n* with independent variables $X_1, X_2, \ldots X_n$ drawn from the population with mean μ and variance σ^2, then the sample mean $\bar{X}$ has an expected value μ and variance $\frac{\sigma^2}{n}$
- We also know from the **Central Limit Theorem** that **the sampling distribution** of $\bar{X}$ is approximately normal for sufficiently large n, even when $X_1, X_2, \ldots X_n$ (sample variables) are not normally distributed.
- With the sampling distribution of the sample mean known, how can we draw inference about the population mean?

Normal Probability for Sample Mean

Now, we know that if $Z \sim N(0,1)$, then

$$P\left(-1.96 < Z < 1.96\right) = 0.95$$

And since $\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$, we have that

$$P\left(-1.96 < \frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) = 0.95$$

Thus,

$$P\left(\mu - 1.96\frac{\sigma}{\sqrt{n}} < \bar{X} < \mu + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Rearranging terms, we have

$$P\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} < \mu < \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

Confidence Intervals for Mean

In other words, there is 95% probability that the *random interval*

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

will cover μ.

Calculating a 95% Confidence Interval

For the time being, we'll continue to assume that σ is known. To calculate a 95% **confidence interval** for the population mean μ

1. Take a random sample of size n and calculate the sample mean $\bar{X}$.
2. If n is large enough, $\bar{X} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right)$ (by the CLT).
3. The confidence interval is given by

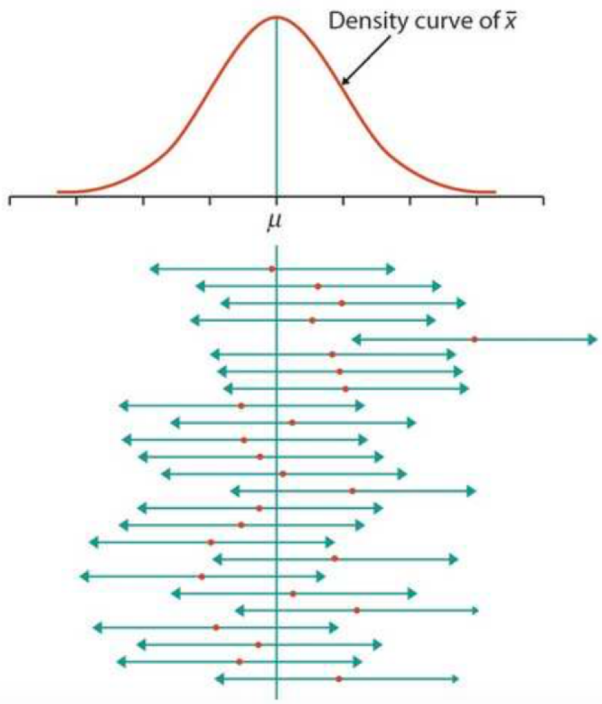
$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \quad \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

Interpretation of Confidence Intervals

Suppose we repeat the following procedure multiple times:

1. Draw a random sample of size n
2. Calculate a 95% confidence interval for the sample

95% of the intervals thus constructed will cover the true (unknown) population mean.



Interpretation of Confidence Intervals (CI): Example

- Consider estimating the speed of light using 64 measurements with sample mean $\bar{x}$ = 298054 km/s.
- Assume we know (from previous experience) that the SD of measurements made using the same procedure is 60 km/s.
- What is a 95% CI for the true speed of light?
- **Incorrect:**
  - There is a 95% probability that the true speed of light lies in the interval (298039.3, 298068.7).
  - In 95% of all possible samples, the true speed of light lies in the interval (298039.3, 298068.7).

Interpretation of Confidence Intervals: Example

- What is a 95% CI for the true speed of light?
- **Correct:**
  - There is 95% confidence that the true speed of light lies in the interval (298039.3, 298068.7).
  - There is 95% probability that the true speed of light lies in the random interval

$$\left(\bar{X} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{X} + 1.96\frac{\sigma}{\sqrt{n}}\right)$$

  - If we repeatedly draw samples and calculate confidence intervals using this procedure, 95% of these intervals will cover the true speed of light.

General Form of Confidence Intervals

In general, a CI for a parameter has the form

$$\text{estimate} \pm \text{margin of error}$$

where the margin of error is determined by the confidence level $(1-\alpha)$, the population SD σ, and the sample size n.

A $(1-\alpha)$ confidence interval for a parameter θ is an interval computed from a SRS by a method with probability $(1-\alpha)$ of containing the true θ.

For a random sample of size n drawn from a population of unknown mean μ and known SD σ, a $(1-\alpha)$ CI for μ is

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

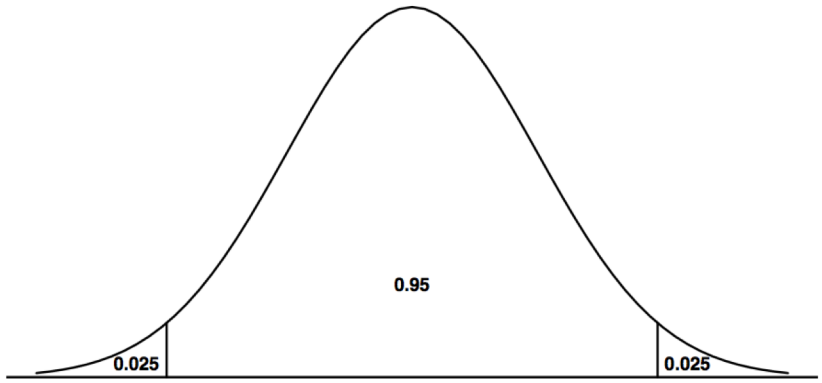
Here z^* is the **critical value**, selected so that a standard Normal density has area $(1-\alpha)$ between $-z^*$ and z^*. The quantity $z^*\sigma/\sqrt{n}$, then, is the **margin error**.

If the population distribution is normal, the interval is *exact*. Otherwise, it is *approximately correct for large* n.

Finding Critical Values for CI

For a given confidence level $(1-\alpha)$, how do we find z^*?

Let $Z \sim N(0,1)$:



$$P(-z^* \leq Z \leq z^*) = (1-\alpha) \iff P(Z < -z^*) = \frac{\alpha}{2}$$

Thus, for a given confidence level $(1-\alpha)$, we can look up the corresponding z^* value on the Normal table.

Common z^* values:

Confidence Level	90	95	99
z^*	1.645	1.96	2.576

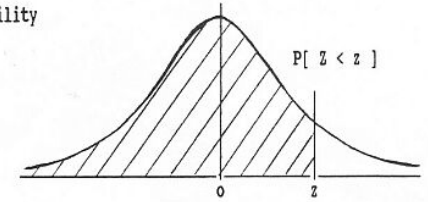
Standard Normal Table

STANDARD STATISTICAL TABLES

1. Areas under the Normal Distribution

The table gives the cumulative probability up to the standardised normal value z

i.e. $$P[\ Z < z\ ] = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} \exp(-\tfrac{1}{2}Z^2)\, dZ$$



z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7854
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8804	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9874	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9924	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
z	3.00	3.10	3.20	3.30	3.40	3.50	3.60	3.70	3.80	3.90
P	0.9986	0.9990	0.9993	0.9995	0.9997	0.9998	0.9998	0.9999	0.9999	1.0000

Confidence Interval: Population Mean

- A 100(1-α) % confidence interval for μ is

$$\left(\overline{X} - z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}}, \quad \overline{X} + z_{1-\alpha/2} \times \frac{\sigma}{\sqrt{n}} \right)$$

- This formula requires knowledge of the population variance (σ^2). If we knew the population variance, this is the formula we would use. In practice, we do not know the population variance.
  - So we use a slightly different formula

Student's *t* Distribution

If the true population SD, σ, is unknown, we estimate σ using the sample SD S.

When the standard deviation of a statistic is estimated from the data, the result is called the standard error of the statistic. The standard error of the sample mean $\bar{x}$ is $SE_{\bar{X}} = \frac{S}{\sqrt{n}}$.

Now, instead of dealing with

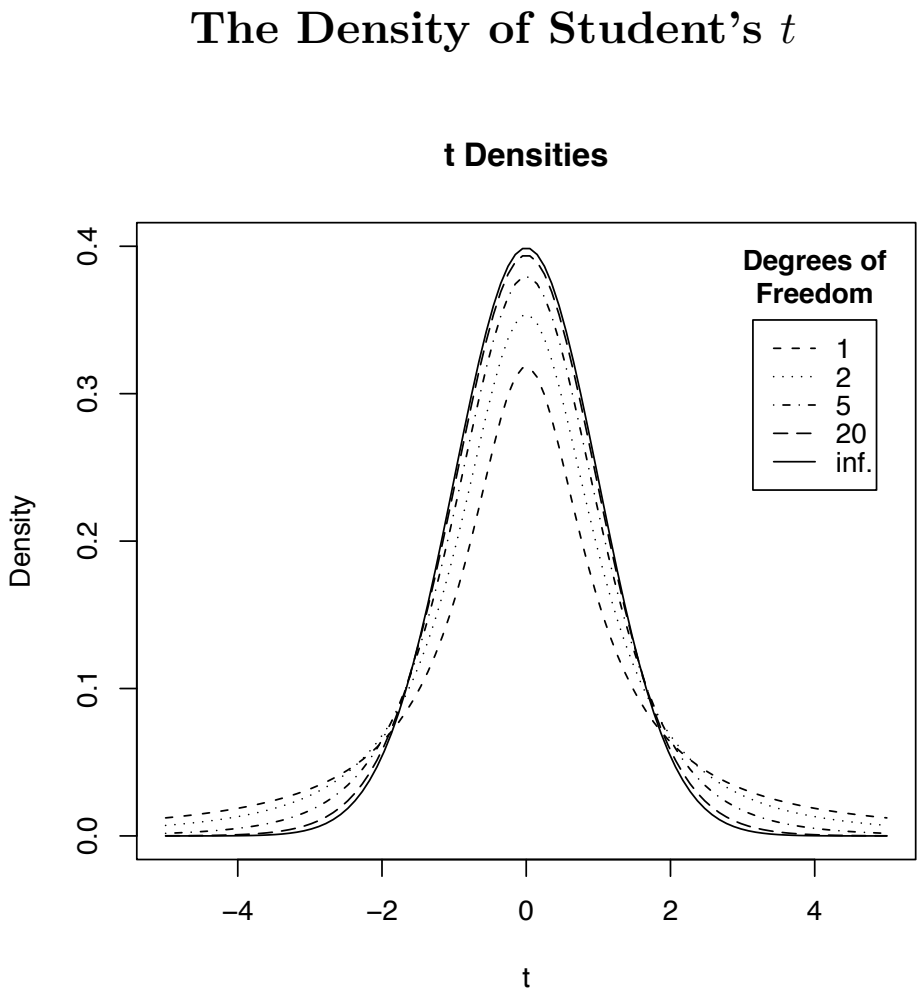
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

we are interested in the quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

Here, $t_{(n-1)}$ is Student's *t* distribution, with $n - 1$ *degrees of freedom.*

Density of Student's *t* Distribution



Properties of Student's *t* Distribution

- Symmetric about zero
- Bell-shaped, similar to normal distribution
- More spread out than normal, i.e., heavier tails than a normal
- Exact shape depends on the degrees of freedom
- As the number of degrees of freedom (e.g., the sample size) increases, the *t* distribution converges to the Normal distribution.

Confidence Intervals with Unknown σ

Recall that, for a population with unknown μ and *known* σ, $100(1-\alpha)\%$ CI for μ, based on an SRS $x_1, x_2, \ldots, x_n$ is given by

$$\left(\bar{x} - z^* \frac{\sigma}{\sqrt{n}},\ \bar{x} + z^* \frac{\sigma}{\sqrt{n}}\right)$$

If σ is unknown,

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim t_{(n-1)}$$

so we substitute a t-critical value, t^* for z^*:

$$\left(\bar{x} - t^* \frac{s}{\sqrt{n}},\ \bar{x} + t^* \frac{s}{\sqrt{n}}\right)$$

The critical value, t^*, is chosen such that $100(1-\alpha)\%$ of the area under the $t_{(n-1)}$ density lies between $-t^*$ and t^*.

The area to the right of t^* should be $\frac{\alpha}{2}$ and the area to the left of $-t^*$ should be $\frac{\alpha}{2}$.

What would t^* be for a 95% confidence interval for a t distribution with 20 degrees of freedom, i.e. $n = 21$?

R: t-distribution Quantiles

- To obtain quantiles/percentiles for a t distribution in R, you only need to use the **qt()** function. Use the help function in R or more details: ***?qt***
- For examples, to obtain the 95th percentile for a t distribution with 20 degrees of freedom, the following command can be used:

```
> qt(p=.95,df=20)
[1] 1.724718
```

- To obtain the 97.5th percentile (or critical value t* for a 95% confidence interval) for a t distribution with 1000 degrees of freedom, the following command can be used:

```
> qt(p=.975,df=1000)
[1] 1.962339
```

  - Note that the 97.5th percentile for a normal distribution is 1.96

t-based critical values

df (n-1)	Critical value for 95% CI
10	2.22
20	2.09
50	2.01
100	1.98
200	1.97
300	1.967
⋮	⋮
Normal	1.96

t-distribution and confidence Intervals

- Whenever we make a confidence interval for the mean of a continuous variable, we must also estimate the population variance.
- This implies we should take our “critical value” from a t-distribution instead of the standard Normal distribution.
  - E.g., for a 95% confidence interval, the critical value of a standard Normal distribution is 1.96. The corresponding value from a t-distribution will be a bit larger.
  - As seen on the previous slide, the critical values for the t-distribution are very close to Normal for large degrees of freedom, so using the normal critical values will generally be fine in practice for large sample sizes (e.g., 200 or more).

Confidence Intervals

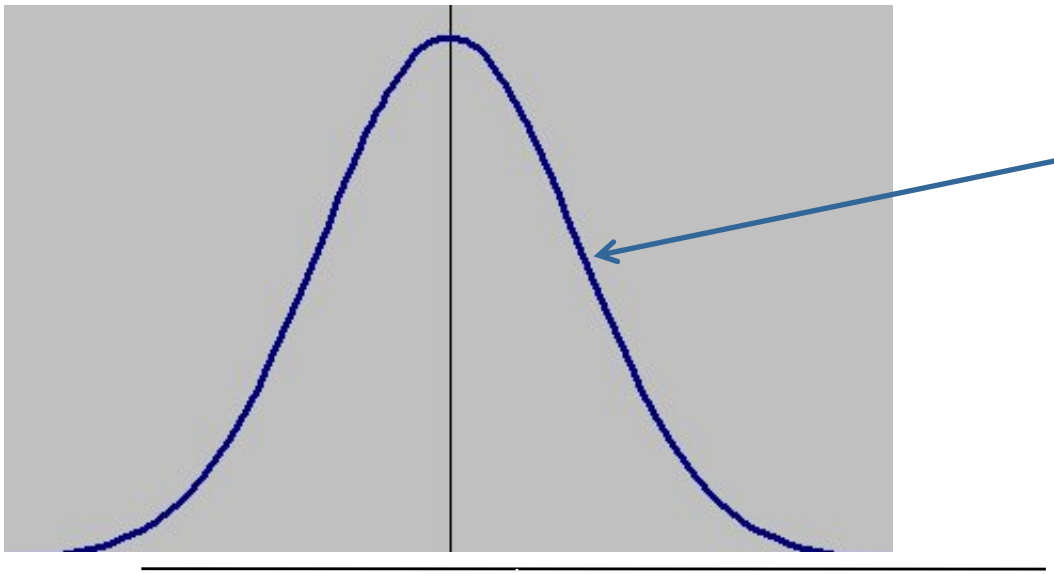
- The most common confidence intervals reported are 95% confidence intervals followed by 99% confidence intervals (distant second).

Confidence Intervals: Thought Exercise

We want to estimate the mean LDL cholesterol level among senior citizens. The mean in our sample of 50 senior citizens is 132 mg/dL.

Ask: If the true mean in the population were some # mg/dL, would a sample mean of 132 be surprising?

Confidence Intervals: Thought Exercise



sampling distribution of the sample mean for a sample of size 50 when the population mean is 130

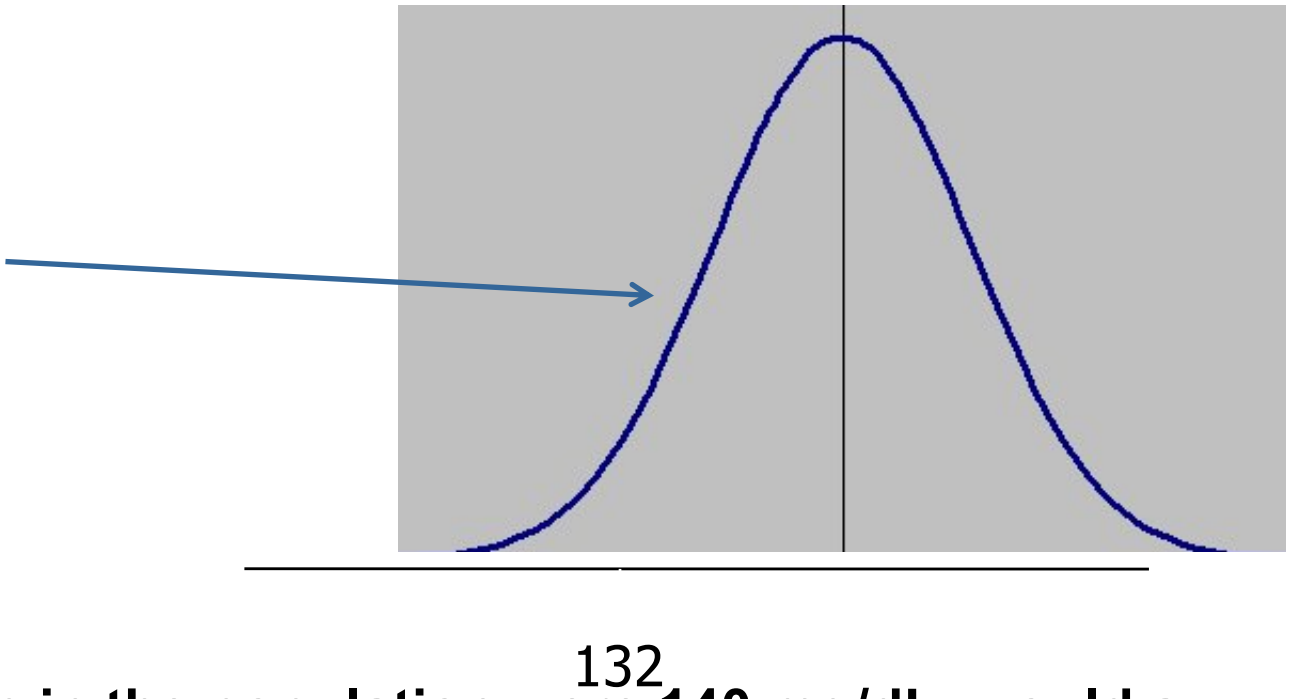
132

Ask: If the true mean in the population were 130 mg/dL, would a sample mean of 132 be surprising?

Ans: No. If the true mean in the population were 130 mg/dL, a sample mean of 132 would not be surprising.

Confidence Intervals: Thought Exercise

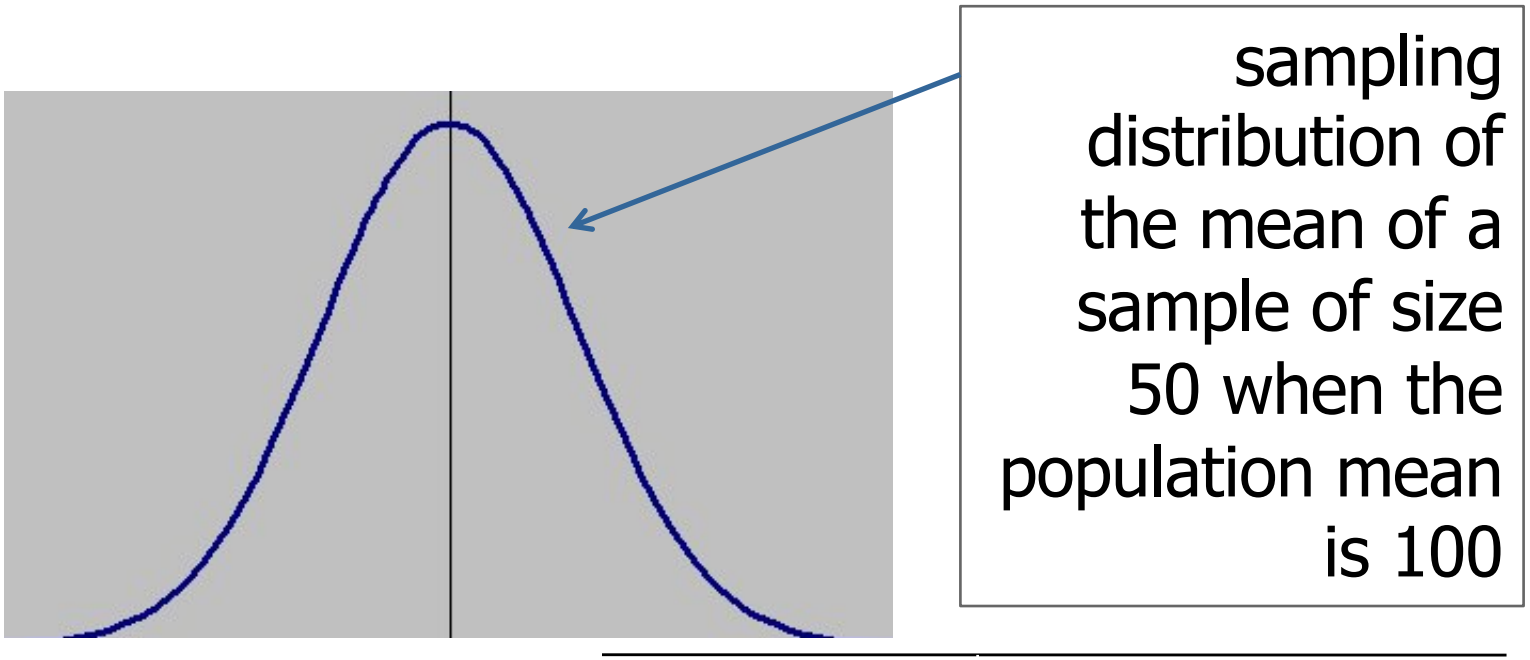
sampling distribution of the sample mean for a sample of size 50 when the population mean is 140



Ask: If the true mean in the population were 140 mg/dL, would a sample mean of 132 be surprising?

Ans: No. If the true mean in the population were 140 mg/dL, a sample mean of 132 would not be surprising.

Confidence Intervals: Thought Exercise



Ask: If the true mean in the population were 100 mg/dL, would a sample mean of 132 be surprising?

Ans: Yes. If the true mean in the population were 100 mg/dL, a sample mean of 132 would be surprising.

Confidence Intervals: Reporting

- The mean cholesterol level in our sample of 50 senior citizens is 132 mg/dL. Suppose that the margin of error for a 95% confidence interval is 10.
- In this example, we might report the following:
  - "With 95% confidence, the mean cholesterol level among senior citizens is between 122 and 142"
  - Or, "The data are consistent* with a mean cholesterol level among senior citizens between 122 and 142"
  - Or, "Our data would not be unusual if the mean cholesterol level among senior citizens is between 122 and 142"

Confidence Intervals: Incorrect Interpretation

Wrong: "There is a 95% chance that mean cholesterol level among senior citizens is between 122 and 142"

Wrong: "We estimate that 95% of senior citizens have cholesterol level between 122 and 142"

Wrong: "If we repeated the study, there is a 95% chance that we would get a mean cholesterol between 122 and 142"

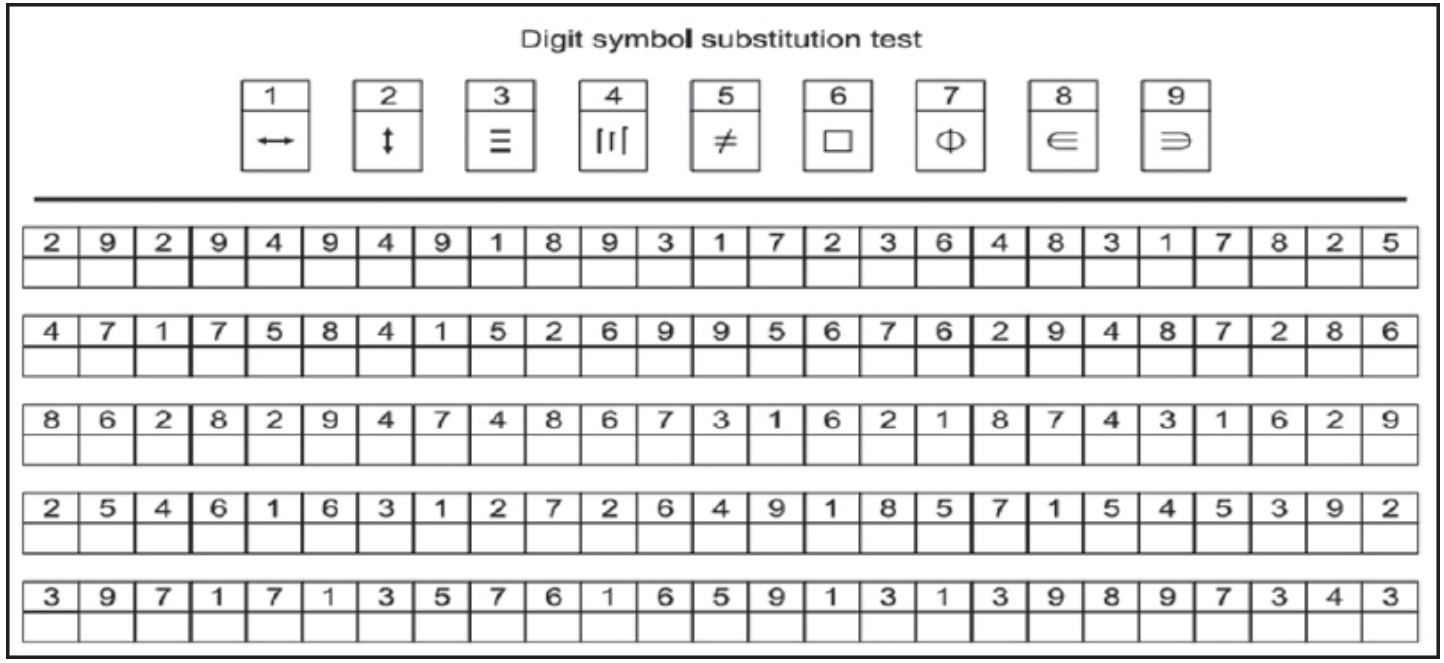
Example: Inference of Cognitive Function in Elderly Adults

Example: Cognitive Function in Elderly Women

- Cardiovascular Health Study of elderly adults aged 65 years and older
- Data was collected at baseline on study participants for various behavioral (e.g., smoking, alcohol consumption), and functional (e.g., ability to perform routine tasks) measures
- Digit symbol substitution test (DSST – a test of attention), and aging was measured at baseline for 3,542 subjects

Example: Mental Function

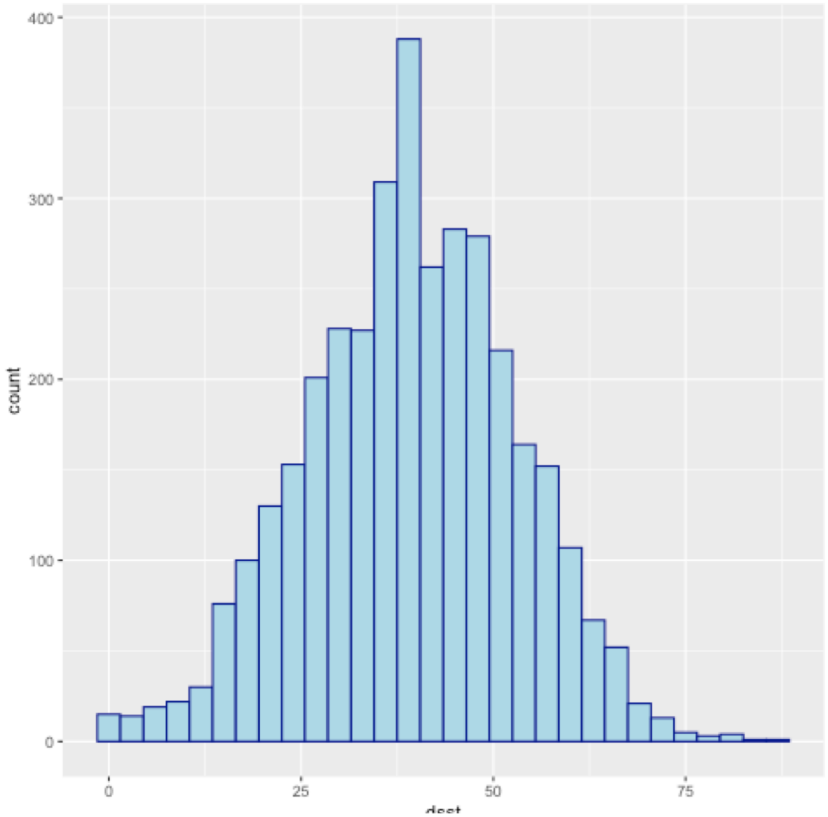
- DSST is a neuropsychological test
  - Consists of a list of digit-symbol pairs.
  - Under each digit the subject should write down the corresponding symbol as fast as possible.
  - The number of correct symbols within the allowed time (e.g. 2 minutes) is measured.



Example: Cognitive Function in Elderly Women

- Suppose we are interested in inference about the mean DSST score in a population of similar elderly adults
- We can obtain a point estimate and confidence interval for the population mean DSST score

Example: DSST




```
> summary(dsstdata$dsst)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
   0.00   30.00   40.00   39.36   49.00   87.00
> sd(dsstdata$dsst,na.rm=TRUE)
[1] 13.59611
```

Example: Inference on Cognitive Function in Elderly Adults

$$\bar{x} = 39.36, \quad s = 13.596, \quad n = 3542$$

- The point estimate for the mean DSST score in a population of elderly adults aged 65 and older is the sample mean, which is 39.36.
- A 95% confidence interval for mean DSST score in this population is:

$$\left(39.36 - 1.96\frac{13.596}{\sqrt{3542}}, 39.36 + 1.96\frac{13.596}{\sqrt{3542}}\right)$$

$$= \left(39.36 - 0.448, 39.36 + 0.448\right) = (38.91, 39.81)$$

- So with 95% confidence, the mean DSST score in a population of similar elderly adults will be between 38.91 and 39.81.

Comments on the Example: Scientific Reporting

- “In a sample of 3,542 elderly adults, the mean DSST score was 39.4 with a standard deviation of 13.6. With 95% confidence, the observed data are consistent with a mean DSST score between 38.91 and 39.81 for a population of similar elderly adults.”

Summary: Inferential Statistics

- Summary measures on populations are called *population parameters* or *parameters*. Population parameters are unknown numbers.
- Inferential statistics computed on samples are used to estimate population parameters
- We don't expect to learn the exact value of a population parameter. We use statistical theory to quantify the uncertainty of our estimate.
  - Statistical theory tells us about the sampling distribution of our statistic, even though we only actually observe one value from the sampling distribution.

Summary: Inferential Statistics

- In this lecture we introduced
  - Point Estimates
  - Interval Estimates (Confidence Intervals)
- A point estimate is our single "best informed guess" of the value of the population parameters. Point estimates by themselves are not very useful. We know there is uncertainty.
- A point estimate accompanied by a confidence interval *acknowledges* the uncertainty in a point estimate and *quantifies the degree* of that uncertainty.