5 Estimation

Simple random sample

A simple random sample (SRS) of size $n$ is a sample chosen so that each subset of $n$ individuals is _______. To draw a simple random sample of size $n$ from a population of size $N$,

- number individuals in population with 1 through $N$
- generate $n$ distinct random integers in ________, and use the corresponding individuals

Each sample in STAT 371 is a simple random sample. (Was the sample in the article you just read an SRS? If not, the conclusion may be ________.)

A note on independence in an SRS

Many theorems require the independence of items in a random sample. Two random variables independent if the realization of either one does not change the distribution of the other. e.g.

- Independent RVs:
- Dependent RVs:

An SRS is drawn ________, that is, an item is not replaced in the population after it is selected (so it cannot be selected more than once). Items in an SRS are not independent.

e.g. Put ten balls labeled 0 through 9 in a bucket.
P(draw 3) = __________

Suppose we draw a 3; then P(draw 3) = __________

On the other hand, to sample ________, replace an item after selecting it (so it may get selected more than once). Items sampled with replacement are independent.

e.g. When sampling with replacement, even after drawing 3, P(draw 3) = __________.

e.g. For a large population, the difference is negligible: with 10000 each of the ten balls in a bucket, drawing a 3 changes P(draw 3) from ________ to _________. We’ll treat items in a SRS as independent, provided the sample size is ________ relative to the population size (which it should always be in this course).

A collection of random variables $X_1, \ldots, X_n$ are independent and identically distributed (IID) if they are all independent and all have the same distribution. We’ll suppose that each ________ is the realization of IID $X_1, \ldots, X_n$. 
Estimating a population mean, $\mu$

e.g. A car manufacturer uses an automatic device to paint engine blocks. Since engine blocks get very hot, the paint must be heat-resistant and of a minimum thickness. A warehouse contains thousands of painted blocks. The manufacturer wants to know the average amount of paint applied, so 16 blocks are selected at random, and the paint thickness is measured in mil ($\frac{1}{1000}$ inch), with these results:

1.29, 1.12, 0.88, 1.65, 1.48, 1.59, 1.04, 0.83, 1.76, 1.31, 0.88, 1.71, 1.83, 1.09, 1.62, 1.49

Before sampling, we regard $X_1, \ldots, X_{16}$ as ___________ with ___________ mean $\mu$ and ___________ variance $\sigma^2$.

How should we estimate $\mu$? Estimator $\hat{\mu} =$ _________________.

Note that an ___________ is the formula that describes how the sample will be used to compute a guess about $\mu$; it’s a random variable. The number computed from the actual sample data is an ___________, a realization of that RV. Our estimate is _________________.

| Theorem: If $X_1, \ldots, X_n$ are IID with mean $\mu$ and variance $\sigma^2$, then $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$ has mean $E(\bar{X}) =$ ___________ and variance $\text{VAR}(\bar{X}) =$ ___________. |

Proof:

The standard error of an estimator is its estimated standard deviation. e.g. The true standard deviation of $\bar{X}$ is ___________; the estimated standard deviation of $\bar{X}$ is ___________, which is its standard error. e.g. The standard error of the mean for the paint data is ___________.

A point estimate alone isn’t very useful. Reporting it with its standard error is useful, but it’s more common to report a ________________ around a point estimate: coming next.

Did our sample come from a normal distribution?

In many common situations, it is reasonable to assume that our sample is from a ___________ population. This leads to a strong statement about the distribution of the sample mean:

| Theorem: If $X_1, \ldots, X_n$ is a simple random sample from a normal population with mean $\mu$ and variance $\sigma^2$ (so $X_i \sim N(\mu, \sigma^2)$), then $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$. |

Soon we will use this theorem to derive a confidence interval. First, let’s look at one way to assess whether a particular sample came from a normal population.
Normal probability plots

Many textbooks and statisticians use a normal probability plot (or normal quantile-quantile plot or normal QQ plot) to decide whether a data set is plausibly a simple random sample of size $n$ from a normal distribution. This plot depicts the $\frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n}{n}$ quantiles from $N(0, 1^2)$ on the $x$-axis against the sorted data set ($\approx$ the corresponding quantiles of the population from which the sample was drawn) on the $y$-axis. The idea is that, if the points more-or-less line up, the data are plausibly from a normal distribution. If the points do not line up, the data are not plausibly from a normal distribution. Here are some details:

For large samples, this seems \underline{plausibly}. For small samples, I'm \underline{cautious}, as I can't tell the difference between \underline{plausibility} in sampling and non-normality in the population.

e.g. In R, try $n = 1000; x = \text{rnorm}(n); \text{qqnorm}(x)$. Then try $n = 10$ or $n = 30$ many times. Also try replacing $\text{rnorm}$ with $\text{runif}$ (thin tails, uniform(0, 1)) and $\text{rlnorm}$ (right-skewed, exp($N(0, 1)$)).
The Central Limit Theorem

The Central Limit Theorem (CLT) says that the mean, $\bar{X}$, of a large enough sample from (almost) any distribution with finite $\mu$ and $\sigma$, is $\approx$ __________:

If $X_1, \cdots, X_n$ is a simple random sample from a population with mean $\mu$ and variance $\sigma^2$, and $n$ is __________, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ (approximately).

($n > 30$ often counts as “large enough”).

e.g. Here is a simulation of the generation of many random samples from the discrete distribution with mass function $p(x) = \frac{1}{10}$ for $x \in \{0, 1, \cdots, 9\}$ (and 0 otherwise):

![Histograms of sample means](image)

e.g. An insurance company knows that in the population of millions of homeowners, the mean annual loss from fire is $\mu = $ $250$ and the standard deviation is $\sigma = $ $1000$. (The loss distribution is strongly right-skewed, since most policies have no loss but a few have large losses.) If the company sells 10,000 policies, can it safely base its rates on the assumption that the average loss will be no greater than $\$275$?
Confidence Intervals for an Unknown Population Mean \( \mu \)

We have two situations in which \( \bar{X} \sim N(\mu, \sigma^2/n) \):
(1) the population is \( N(\mu, \sigma^2) \), for small or large \( n \); then “\( \sim \)”, above, is exact. (2) The sample size \( n \) is large enough that CLT applies: then “\( \sim \)” above is approximate. Suppose, then, that \( \bar{X} \sim N(\mu, \sigma^2/n) \).

Here we construct an interval around \( \bar{X} \) which contains \( \mu \) for a proportion \( 1 - \alpha \) of random samples, where \( \alpha \in (0, 1) \). \( 100\%(1 - \alpha) \) is the confidence level of the interval.

Let \( z_{\alpha/2} \) = the \( z \)-score cutting off a right tail area of _____ from \( N(0, 1) \) (draw).

E.g. For the conventional confidence level 95%, \( \alpha = \) _______ and \( z_{\alpha/2} = \) __________

Then \( P(-z_{\alpha/2} < Z < z_{\alpha/2}) = 1 - \alpha \) (draw). Unstandardize using \( Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \) (where \( \sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \)) to get

\[ P(-z_{\alpha/2} < \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2}) = 1 - \alpha, \]

which we solve in two ways:

- for \( \bar{X} \) in the middle: \( P(\mu - z_{\alpha/2}\sigma/\sqrt{n} < \bar{X} < \mu + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha \) (see picture)
- for \( \mu \): \( P(\bar{X} - z_{\alpha/2}\sigma/\sqrt{n} < \mu < \bar{X} + z_{\alpha/2}\sigma/\sqrt{n}) = 1 - \alpha \) (see picture)

For this \( \bar{x} \), \( \mu \) is _________ the confidence interval. This happens with probability ______.

\( \bar{X} \sim N(\mu, \sigma^2/n) \)

For this \( x \), \( \mu \) is _________ the confidence interval. This happens with probability ______.
That is, \( \bar{X} \pm z_{\alpha/2} \sigma_{\bar{X}} = \bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \) contains \( \mu \) for a proportion \( 1 - \alpha \) of random samples. It’s the \( 100%(1 - \alpha) \) confidence interval for \( \mu \). This form is useful when we know \( \sigma \), which is ______.

e.g. Suppose we know \( \sigma_{\text{paint thickness}} = 0.30 \) micrometers. Find a 95% CI for \( \mu \).

\( n = \) ______ ; Is \( n \) large enough or is sample from normal population? (Try ```qqnorm(paint)```.)

\[
1 - \alpha = \text{__________} \quad \Rightarrow \quad \alpha = \text{__________} \quad \Rightarrow \quad z_{\alpha/2} = \text{__________}
\]

\( \bar{x} = \text{__________} \), error margin = \( \text{________________________} \)

\( \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} = \text{________________________} \)

With what probability does our interval contain \( \mu \)? \( \text{__________} \)

**How Confidence Intervals Behave**

- \( \bar{X} \pm \frac{\sigma}{\sqrt{n}} \) is a 68% confidence interval for \( \mu \)
- \( \bar{X} \pm 1.96 \frac{\sigma}{\sqrt{n}} \) is a 95% confidence interval for \( \mu \) (and \( 1.96 \approx \) ______)
- \( \bar{X} \pm \frac{\sigma}{\sqrt{n}} \) is a 99% confidence interval for \( \mu \)
- \( \bar{X} \pm \frac{\sigma}{\sqrt{n}} \) is a 99.7% confidence interval for \( \mu \)

We want high confidence and a small margin of error, but the margin is \( z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \), which gets smaller when \( z_{\alpha/2} \) gets smaller, which corresponds to \( (1 - \alpha) \) getting smaller too. Extreme cases are that we can have confidence approaching 100% as the margin approaches \( \text{__________} \), or we can have confidence approaching \( \text{__________} \) as the margin approaches 0.

**Choosing the Sample Size**

Good news is that the margin also gets smaller as \( \text{_____________} \). For a desired margin of error \( m \), we can find the required sample size:

\[
m = z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \Rightarrow \quad (\text{Use } \sigma \approx s \text{ in the usual case where we don’t know } \sigma.)
\]
The Student’s $t$ Distribution

Now suppose we don’t know $\sigma$. Define the random variable $T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$. $T$’s distribution isn’t normal; it’s the Student’s $t$ distribution with $n - 1$ degrees of freedom, denoted $t_{n-1}$. (“Student” is a pseudonym for William Gosset, a statistician at ___________.) Here are some of its properties:

- $T$ is a sample version of a ___________, estimating how far $\bar{X}$ is from ________________, in ________________

- $t_{n-1}$ looks like $N(0, 1)$: symmetric about ____, ___________-peaked, and _______-shaped

- $T$’s variance is __________ than $Z$’s because estimating $\sigma$ (______) by $S$ (________) gives $T$ more variation than $Z$: $t_{n-1}$ is shorter with thicker tails (draw $N(0, 1)$ and $t_{6-1}$)

- As $n$ increases, $t_{n-1}$ gets closer to __________ $(S$ becomes a ________________ of $\sigma)$; in the limit as $n \to \infty$, they’re __________

Let $t_{n-1,\alpha} =$ the critical value $t$ cutting off a ________________ area of $\alpha$ from $t_{n-1}$ (draw). The posted Student’s $t$ table gives __________ tail probabilities, using $\nu$ (“nu”) for $n - 1$.

![Diagram showing $T \sim t_{n-1}$]

$\alpha =$ shaded area

0 $t_{n-1,\alpha}$

e.g. Use the $t$ table to find the critical value $t$

- cutting off a right tail area of .05 from the $t_{6-1}$ distribution: $t_{5,.05} =$ __________
- such that the area under the $t_{22-1}$ curve between $-t$ and $t$ is 98%
- such that the area under the $t_{25-1}$ curve left of $t$ is .025
Confidence Intervals Using the Student’s $t$ Distribution

Theorem: If $X_1, \ldots, X_n$ is a simple random sample from a normal population with mean $\mu$ and variance $\sigma^2$ (so $X_i \sim N(\mu, \sigma^2)$), or $n$ is large enough, then $\bar{X} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}}$ contains $\mu$ for a proportion $1 - \alpha$ of random samples.

Proof:

This is the $100\%(1 - \alpha)$ confidence interval for $\mu$, useful when we don’t know $\sigma$ and have a sample of any size from a normal population or a large sample from (almost) any population.

e.g. Make a 95% confidence interval for the paint data (supposing now that we don’t know $\sigma$).

$n =$ ______ ; Is $n$ large enough or is sample from normal population? (Try \texttt{qqnorm(paint)}.)

$1 - \alpha =$ __________ $\Rightarrow \alpha =$ __________ $\Rightarrow t_{n-1,\alpha/2} =$ __________

$\bar{x} =$ __________

$s =$ __________

error margin = ________________

$\bar{x} \pm t_{n-1,\alpha/2} \frac{s}{\sqrt{n}} =$ ________________

With what probability does our interval contain $\mu$? __________

Compare this unknown-$\sigma$ interval to our earlier known-$\sigma$ version.

Example

e.g. In a sample of 100 boxes of a certain type, the average compressive strength was 6230 N, and the standard deviation was 221 N.

a. Find a 95% confidence interval for the mean compressive strength.

b. Find a 99% confidence interval for the mean compressive strength.
Bootstrap Methods

So far, our discussion of estimating the population mean $\mu$ has assumed either the population is normal, so that $\bar{X}$ is also __________, or the sample size is __________ for the CLT to indicate that $\bar{X}$ is approximately normal. What if neither is true?

e.g. Secondhand smoke presents health risks, especially to children. A SRS was taken of 15 children exposed to secondhand smoke, and the amount of cotanine (a metabolite of nicotine) in their urine was measured. Cotanine in unexposed children should be below 75 units. The data were:

29, 30, 53, 75, 34, 21, 12, 58, 117, 119, 115, 134, 253, 289, 287

First, check graphs to see whether an assumption of a normal population is __________:

![Secondhand smoke and children (histogram and density plot)](image)

This looks pretty bad, so we worry about a normality assumption. The sample is small, so the CLT may not help. Without a normal $\bar{X}$, the quantity

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

will not have a __________ distribution. The *bootstrap* is a sneaky way to estimate the true distribution of this $T$. It estimates the __________ of a statistic by sampling with replacement from a simple random sample from a population. e.g. Here’s a hand-waving account...
To make a bootstrap confidence interval for an unknown mean \( \mu \),

1. Collect one simple random sample of size \( n \) from the population. Compute the sample mean, \( \bar{x} \), which is an estimate of the population mean, \( \mu \).

2. Draw a random sample of size \( n \), \( x^*_1, x^*_2, \ldots, x^*_n \), from the data. Call these observations \( x^*_1, x^*_2, \ldots, x^*_n \). Some data may appear more than once in this resampling, and some not at all.

3. Compute the \( \bar{x}^* \) and \( s^* \) of the resampled data. Call these \( \bar{x}^* \) and \( s^* \).

4. Compute the statistic \( \hat{t} = \frac{\bar{x}^* - \bar{x}}{s^*/\sqrt{n}} \).

5. Repeat steps 2-4 a large number of times, accumulating many \( \hat{t} \)'s. They approximate the sampling distribution of \( T \).

6. Find the \( \alpha/2 \) and \( 1 - \alpha/2 \) quantiles (or critical values) of the approximate sampling distribution, calling them \( \hat{t}_{(\alpha/2)} \) and \( \hat{t}_{(1-\alpha/2)} \).

7. The bootstrap 100(1 - \( \alpha \))\% confidence interval is \( (\bar{x} + \hat{t}_{(\alpha/2)} \frac{s}{\sqrt{n}}, \bar{x} + \hat{t}_{(1-\alpha/2)} \frac{s}{\sqrt{n}}) \).

For the secondhand smoke data, we find \( \bar{x} = 108.4 \) and \( s = 95.6 \). Bootstrapping 1000 times yields the following approximate distribution of \( t \):

Unlike a \( t \) or normal distribution, this distribution is {
\underline{\underline{\text{symmetric}}}}. The quantiles, from R, are \( \hat{t}_{(\alpha/2)} = -3.56 \) and \( \hat{t}_{(1-\alpha/2)} = 1.86 \) (draw), so the interval is \( (108.4 + (-3.56) \frac{95.6}{\sqrt{15}}, 108.4 + (1.86) \frac{95.6}{\sqrt{15}}) = (20.5, 154.3) \).

This interval is not {
\underline{\underline{\text{symmetric}}} \text{ around the point estimate, because the approximate distribution of the statistic is not symmetric.}}

Note: If we took another 1000 resamples, our interval would change a little.

If we had assumed normality and used the Student’s \( t \)-distribution, our interval would have been \( (55.5, 161.3) \), different than the bootstrap interval.
Here is one way to do this bootstrap using R:

```r
# Create a new function, bootstrap(x, n.boot), having two inputs:
# - x is a data vector
# - n.boot is the desired number of resamples from x
# It returns a vector of n.boot t-hat values.

bootstrap = function(x, n.boot) {
  n = length(x)
  x.bar <- mean(x)
  t.hat <- numeric(n.boot) # create vector of length n.boot zeros
  for(i in 1:n.boot) {
    x.star <- sample(x, size=n, replace=TRUE)
    x.bar.star <- mean(x.star)
    s.star <- sd(x.star)
    t.hat[i] <- (x.bar.star - x.bar) / (s.star / sqrt(n))
  }
  return(t.hat)
}

# Use the bootstrap() function to get an approximate sampling
# distribution of T for the smoke data.
smoke = c(29, 30, 53, 75, 34, 21, 12, 58, 117, 119, 115, 134, 253, 289, 287)
smoke.boot <- bootstrap(smoke, 1000)

# Plot the approximate sampling distribution.
hist(smoke.boot, xlab = "Bootstrap t-hat values",
     main = "Approximate Sampling Distribution of T")

# Find quantiles for a 95% confidence interval.
t.lower <- quantile(smoke.boot, probs=.025) # This is our t_{1 - alpha.2}.
t.upper <- quantile(smoke.boot, probs=.975) # This is our t_{alpha/2}.

# Make the interval.
n = length(smoke)
x.bar = mean(smoke)
s = sd(smoke)
x.bar + t.lower * s / sqrt(n)
x.bar + t.upper * s / sqrt(n)
```
Estimation of a Population Proportion

e.g. An accounting firm has a large list of clients (the population), with an information file on each client. The firm has noticed errors in some files and wishes to know the proportion of files that contain an error. Call the population proportion of files in error $\pi$. An SRS of size $n = 50$ is taken and used to estimate $\pi$. Now the firm will decide whether it is worth the cost to examine and fix all the files. Each file sampled was classified as containing an error (call this 1), or not (call this 0). The results are:

Files with an error: 10; files without errors, 40.

To develop an estimator of $\pi$, recall the binomial distribution: $X \sim Bin(n, \pi)$ is the number of successes in $n$ independent trials, each having two possible outcomes (success and failure), and each having probability $\pi$ of success. We found $E(X) = \mu$, $VAR(X) = \sigma^2$.

Our estimator of the population proportion is the sample proportion $P = \hat{\pi} = \frac{X}{n}$. Here are some of its properties:

- $E(P) = \mu$
- $VAR(P) = \sigma^2$
- $SD(P) = \sqrt{\frac{\pi(1-\pi)}{n}}$

This tells us our estimator $P$ is normally distributed for $\pi$, and gives a measure of precision. As in the discussion of $\bar{X}$, we can estimate the standard deviation by plugging in our estimator of $\pi$:

Standard error of $P$ (estimated standard deviation of $P$) $= SE(P) = s_P = \sqrt{\frac{P(1-P)}{n}}$.

To make a CI for $\pi$, we need the distribution of $P$. Its exact distribution is related to the binomial distribution, but making an exact CI based on this fact is difficult. However, the CLT can help. If $n$ is large enough, the conditions of the CLT are met, because $X = \sum Y_i$ (where $Y_i$ is a Bernoulli trial, either 0 or 1), so $P = \frac{\bar{X}}{n} = \frac{1}{n} \sum Y_i$ is a normal distribution. Thus, for large samples, $P$ is approximately normally distributed.

$$P \sim N \left( \pi, \left[ \frac{\pi(1-\pi)}{n} \right]^2 \right) \approx N \left( \pi, \left[ \frac{\pi(1-\pi)}{n} \right]^2 \right)$$

Here is the interval:
Theorem: If $X$ is the number of successes in a large number $n$ of independent Bernoulli trials, each having probability $\pi$ of success, and $P = \hat{\pi} = \frac{X}{n}$, then an approximate 100%(1 $-$ $\alpha$) confidence interval for $\pi$ is $P \pm z_{\alpha/2} \sqrt{\frac{P(1-P)}{n}}$. (Regarding $n$, a rule of thumb says we need ________________, where $\pi$ can be approximated by $P$).

Proof:

e.g. Find a 95% CI for the unknown proportion $\pi$ of defective files.