Sampling review

Sampling involves taking a subset of size n from a large **population**. We hope to infer **parameters** (such as μ, σ,...) of the population from experience with **random samples** about which definite expectations can be formed.

These expectations are facts about the sampling distribution of a statistic – what you would get by continuing to draw random samples of size n, one after another, from the same population.

mean of sampling distribution The mean of the sampling distribution of means, $\mu_{\bar{x}}$ is $E[\bar{X}] = \mu_{\bar{x}} = \mu$ where μ is the population mean.

variance of sampling distribution The variance of the sampling distribution of means, $\sigma_{\bar{X}}^2$ is
$$E[(\bar{X} - \mu)^2] = \sigma_{\bar{X}}^2 = \frac{\sigma^2}{n} \text{ where } \sigma^2 \text{ is the variance of the population}$$

As an estimator of μ the sample mean $\bar{X}$ is **unbiased** – on average it points to the true population value – and **consistent**, meaning that it gets closer and closer as the sample size ramps up.

df again

It's important to divide by $n - 1$ because the "average residual" $\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2$ is not an unbiased estimator of the population variance. That is,

$$E \left[\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] \neq \sigma^2$$

but

$$E \left[\frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})^2 \right] = \sigma^2$$

Vasishth §3.9 uses simulations to demonstrate how using the wrong denominator in sampling will lead to biased estimates of σ.

Hypotheses

In an experiment, one goes in with a hypothesis. A valuable hypothesis will state something general about the entire population. For instance, on the basis of previous research on American children, we might suppose that British children, as well, have a vocabulary of 24,000 words. Interpret this as a proposal about the *population mean*.

> $\mathcal{H}_0$ British children have a 24000-word vocabulary
> $\mathcal{H}_1$ British children have a vocabulary whose size is some other number than 24000

The "alternative" hypothesis $\mathcal{H}_1$ is the one the scientist seeks to support. $\mathcal{H}_1$ is the "research hypothesis" that motivates the study. The "null" hypothesis, $\mathcal{H}_0$ encompasses all the other ways things could turn out.

To reject the null hypothesis is to observe a sample with properties so wild that it is unlikely the null hypothesis can be the case. Thus, the complement of the null hypothesis, namely the alternative hypothesis (which the researcher favored all along) must be true (Table 1).

1

Investigator's Decision	The Real Situation (unknown to investigator)	
	$\mathcal{H}_0$ is True	$\mathcal{H}_0$ is False
Reject $\mathcal{H}_0$	Investigator makes a Type I error	Investigator makes a correct decision
Do not reject $\mathcal{H}_0$	Investigator makes a correct decision	Investigator makes a Type II error

Table 1: Table of Possible Mistakes (p55 of Johnson)

Vasishth's Height Example from 3.7,3.12

What if we were omniscient, and knew that $\mu = 60, \sigma = 4$. Then it would be easy to judge whether a given sample is surprising or not. For instance, in eleven measurements of peoples' heights,

$$X = \{59, 58, 58, 70, 59, 53, 62, 63, 65, 60, 60\}$$

The mean is

$$\bar{X} = 667/11 = 60.63$$

How wacky is this? Is the one-number summary of these observations near or far from $\mu = 60$? If we convert to standardized scores, we can make a probability statement.

$$
\begin{aligned}
z &= \frac{\text{score} - \text{mean}}{\text{standard deviation}} \\
&= \frac{\bar{X} - \mu}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \\
&= \frac{60.63 - 60}{4/\sqrt{11}} = 0.527
\end{aligned}
$$

One standard deviation of this Z-score amounts to $\frac{4}{\sqrt{11}} \approx 1.2$ inches.

These standardized scores are normally-distributed. Recalling properties of the Normal,

the area under the normal curve between -1 and $+1$ is about 68%
the area under the normal curve between -2 and $+2$ is about 95%

we can expect 95% of such 11-observation samples to have means between $60 + (2 \times 1.2)$ and $60 - (2 \times 1.2)$. So, we've observed a run-of-the-mill sample.

1. Population POP exists with parameters $\mu, \dots$ that are either consistent or inconsistent with $\mathcal{H}_0$

2. Random sample SAMP is taken from POP. We expect much less variability (square root law).

3. Did SAMP fall outside a confidence interval for μ based on POP's sampling distribution, assuming $\mathcal{H}_0$ is true?

   Yes If so, this is surprising. Reject $\mathcal{H}_0$ at a observed significance level reflecting how unlikely it would be to get SAMP if $\mathcal{H}_1$ were not true.

   No If not, the null hypothesis cannot be rejected. $\mathcal{H}_1$ might still be true, but this experiment did not detect it.

Figure 1: Road map for sampling experiment

2

Height example with $\mathcal{H}_0$: $\mu = 64$

What if SAMP was systematically chosen the experimental group of people given a shrinking drug, such that the rest of the population is known to have a mean height of 64 inches? The drug maker would like to know if the shrinking drug has any effect, i.e. if the mean of SAMP is different from the no-effect value 64? How likely is our sample if $\mathcal{H}_0$ is true?

$$
\begin{aligned}
z &= \frac{\bar{X} - \mu_0}{\sigma_{\bar{X}}} = \frac{60.63 - 64}{4/\sqrt{11}} \\
&\approx -2.7
\end{aligned}
$$

The number -2.7 is the value of a **test statistic**. Finding a sample such as SAMP, whose mean is more than 2 SDs below what is expected if $\mathcal{H}_0$ were true, is grounds for rejecting the idea that the shrinking drug has no effect. Au contraire, it seems to shrink people.

Even if we don't know σ

Luckily an analogous statistic that does not require knowledge of σ also has a known shape, the t-distribution.

$$
\begin{aligned}
t &= \frac{\text{statistic} - \text{parameter}}{\text{estimated standard error of statistic}} \\
&= \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \\
&= \frac{60.63 - 64}{4.388/\sqrt{n}} \\
&\approx -2.54
\end{aligned}
$$

The t-value, -2.54, is also a Normally-distributed test statistic. It is less extreme than the Z-score computed above. This should make sense, it's based on an estimate of the sample variance that leads us to expect a bit more variability. Still, it points to the same conclusion: the probability of obtaining a t-value this low is only 1.4%.

```
> pt(-2.54, df = 10)
```

```
[1] 0.01468158
```

We are ready to call such a sample evidence against $\mathcal{H}_0$.

```
> samp <- c(59, 58, 58, 70, 59, 53, 62, 63, 65, 60, 60)
> t.test(samp, mu = 64, alternative = "less")

        One Sample t-test

data:  samp
t = -2.5424, df = 10, p-value = 0.01462
alternative hypothesis: true mean is less than 64
95 percent confidence interval:
    -Inf 63.03431
sample estimates:
mean of x
 60.63636
```

Our result about shrinking drugs is significant at the 0.05 level because $0.014 < 0.05$. The **significance level** α is the minimum tolerable probability of making a type I error. β is the probability of making a type II error, and $1 - \beta$ is known as the **power**. Power can be increased by reducing variability or getting more data.

```
> source(file = "shade.tails.R")
> shade.tails(2.54, tail = "lower", df = 10)
```
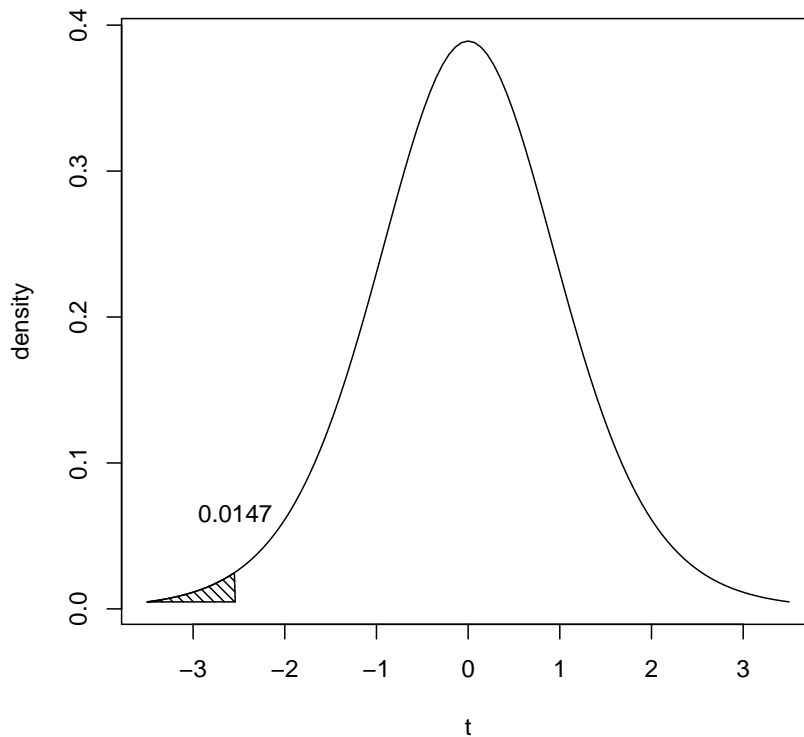

Figure 2: The probability density of t with 10 degrees of freedom. The area of the shaded region at $t \leq -2.54$ indicates that only 1.4% samples are expected to have t-values at or below this level. Diagram created with Keith Johnson's `shade.tails` function.
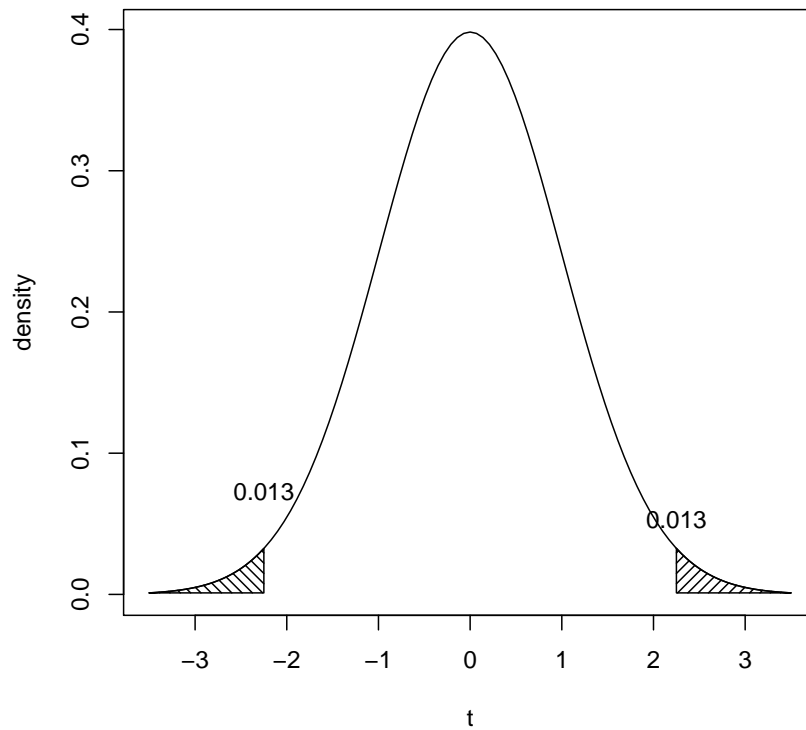
Two-tailed

If $\mathcal{H}_0$ is that British six-year-olds know 24K words, then any evidence of vocabularies bigger *or* smaller than that counts against the null. Observing a $n = 140$ British children with a (sample) mean vocabulary size of $\bar{X} = 24,800$ and sample standard deviation of 4,200,

$$
\begin{aligned}
t & = \frac{\bar{X} - \mu_0}{s/\sqrt{n}} \\
& = \frac{24,800 - 24,000}{4200/\sqrt{140}} \\
& \approx 2.25
\end{aligned}
$$

leads to a t-value that is outside the CI that we would expect if British children really had 24K vocabularies. This sample is evidence against $\mathcal{H}_0$ but the tails are slightly smaller since the 5% probability of receiving a sample with a mean different from 24K has to be shared between the "higher than" and "lower than" possibilities.

```
> source(file = "shade.tails.R")
> shade.tails(2.25, tail = "both", df = 139)
```


Now You Try

1. A sample of 184 children taken an articulation test. Their mean score is 48.8 with standard deviation 12.4. Show that these results are consistent with the null hypothesis that the population mean is $\mu = 50$ against the alternative that $\mu \neq 50$.