Biostats Lecture 7:
Multiple testing; Confidence Intervals

Public Health 783
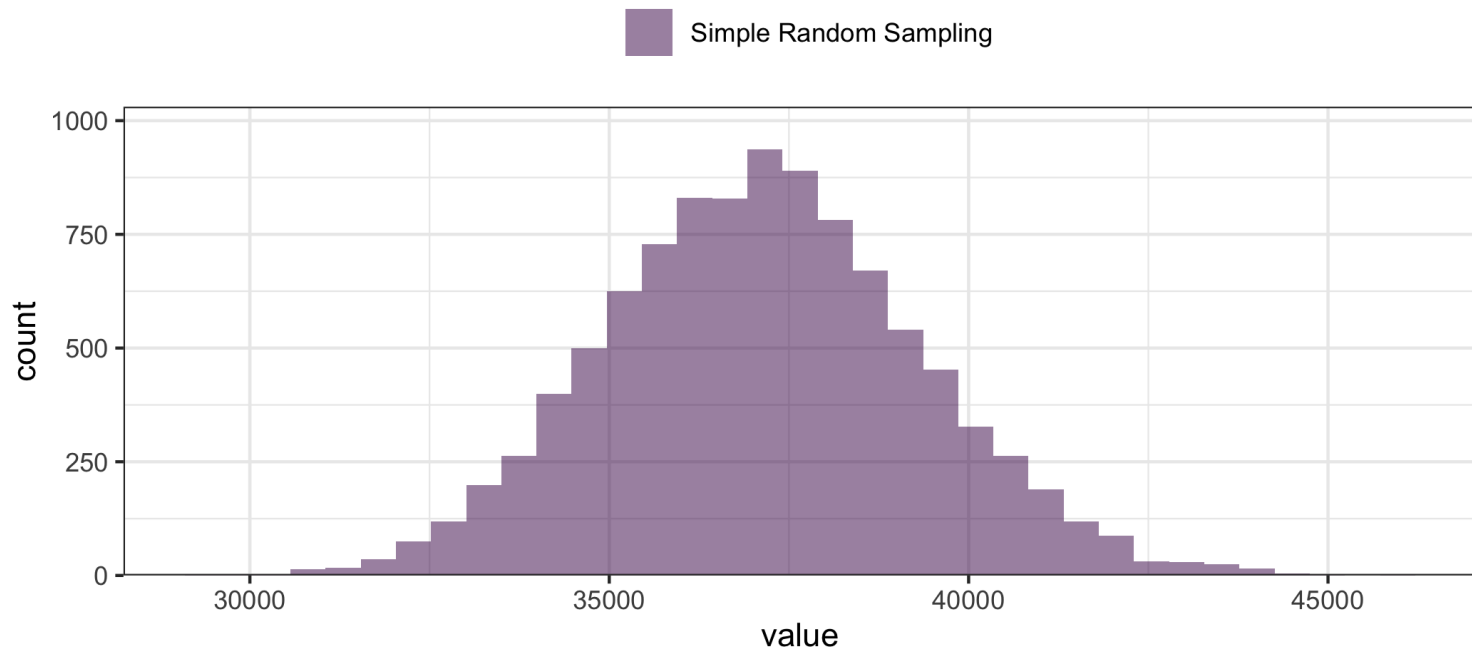
Ralph Trane
University of Wisconsin–Madison

Fall 2019

WISCONSIN
UNIVERSITY OF WISCONSIN–MADISON

But first, sampling bias

To answer, take a simple random sample (i.e. sample individuals at random), calculate the average. What kind of results would we expect?

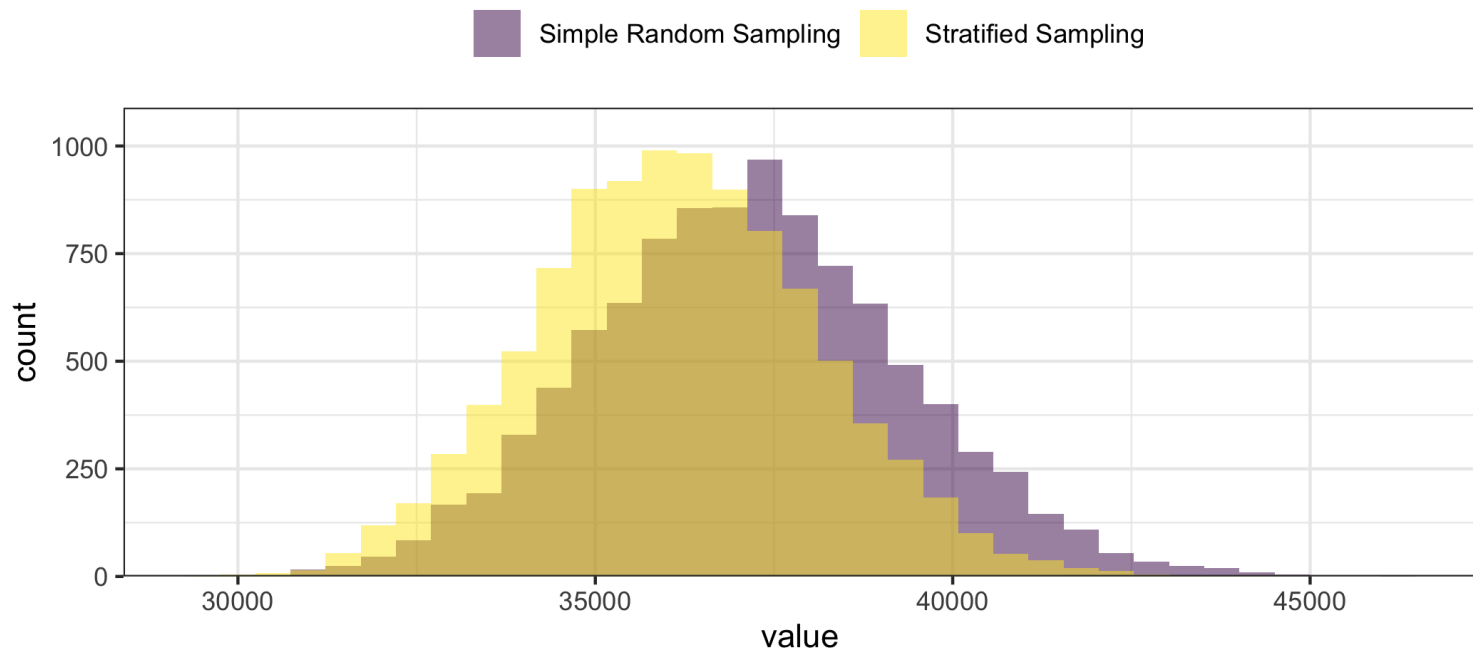Here's the distribution of 10000 samples of size 189.

Would this be ideal, or can we do better?

But first, sampling bias

Some counties are smaller than others, and would not be selected as often when we randomly select people. Should we do stratified sampling? Maybe...

To do so, sample 3 individuals from each of the 63 counties. What would we expect from this stratified sampling scheme, and how does it compare to SRS?
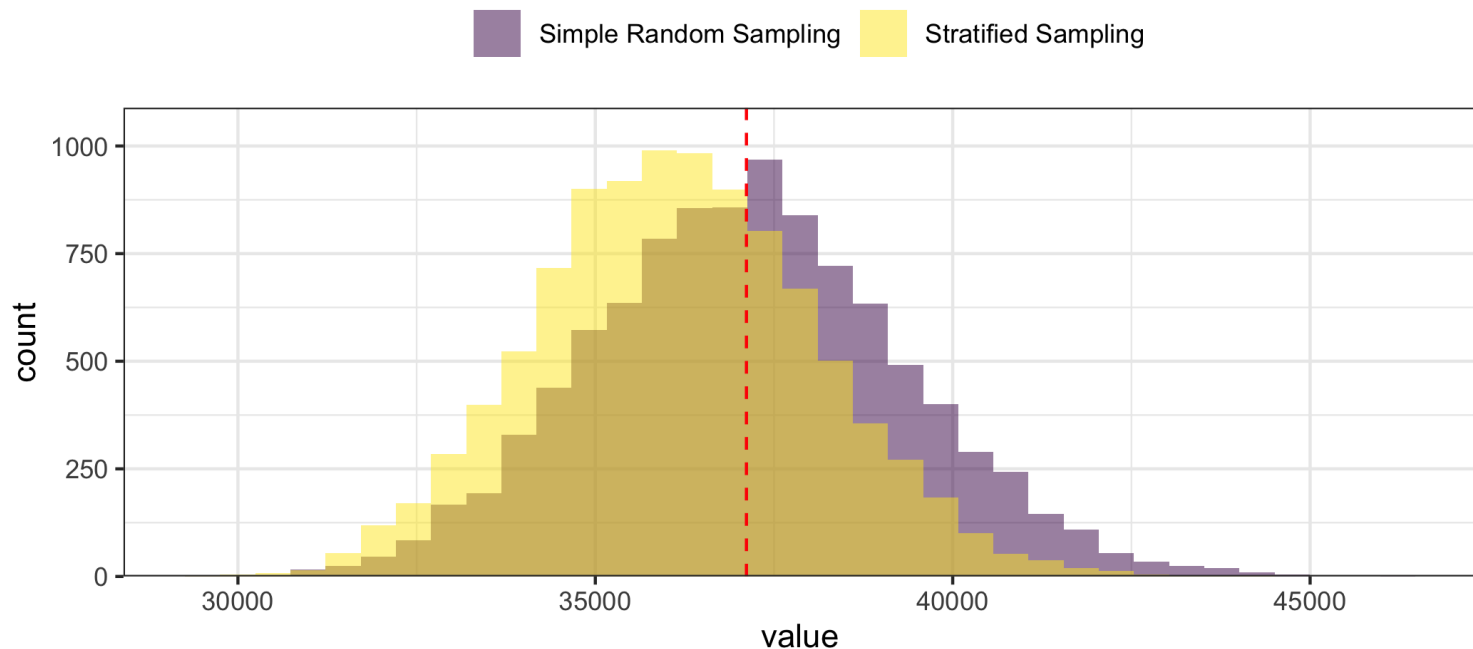
But first, sampling bias

Some counties are smaller than others, and would not be selected as often when we randomly select people. Should we do stratified sampling? Maybe...

To do so, sample 3 individuals from each of the 63 counties. What would we expect from this stratified sampling scheme, and how does it compare to SRS?

Notice how the SRS means are centered around the truth, while the stratified sample means are shifted to the left. Which would we prefer?

But first, sampling bias

Does this mean SRS is always preferable? No. Sometimes stratified is better, *but you have to account for it in your analysis!!*

I.e. a simple average won't be a good estimate of the true population mean. We say that the estimator is *biased* since the expected value is NOT the true value.

However, if done right, stratified sampling (or other sampling schemes) can reduce variation, and therefore give you more confidence in your final estimate.

Recap

Population of interest.

Get a representative sample.

Make assumptions about the way the sampling works (i.e. the distribution of the data).

Set up hypothesis.

Ask: "If the hypothesis is true, what is the probability of seeing data that are this far from the hypothesis?"

If probability is small, reject hypothesis. If probability large, do not reject hypothesis.

(**NOTE**: we NEVER accept the null hypothesis.)

Multiple Testing Problem

Probability of Type I error if we test one hypothesis: 0.05 (or, in general, α)

What if we test more hypotheses?

Assume they're independent, and we test k hypotheses $H_1, H_2, \ldots, H_k$. Then

$$
\begin{aligned}
P(\text{no type I error}) &= P(\text{do not reject any hypotheses}|\text{all hypotheses are true}) \\
&= P(\text{do not reject } H_1 | H_1 \text{ is true}) \cdots P(\text{do not reject } H_k | H_k \text{ is} \\
&= 0.95 \cdots 0.95 = 0.95^k.
\end{aligned}
$$

Probability of rejecting at least one hypothesis *IF* they are all true:
$1 - P(\text{do not reject any hypotheses}|\text{all hypotheses are true})$.

Multiple Testing Problem

Why should we care? Because if we don't, then Jelly Beans Cause Acne!!!

SHOW data has about 109 measured variables. If you go looking for "significant associations", you are bound to find some, even if none are truly present!

Bonferroni Correction

Instead of rejecting when $p < \alpha$, reject when $p < \alpha/m$, where m is the number of tests.

Pro: super simple, and somewhat intuitive

Con: known to be overly conservative (i.e. produces false negatives and thereby reducing power).

Benjamini-Hochberg

A bit more convoluted, but goes as follows:

1. Rank your p-values. Let p_1 denote the smallest, p_2 the second smallest, ..., p_k the k'th smallest, ... p_m the largest,
2. Find the largest k such that $p_k \leq \frac{k}{m}\alpha$,
3. Reject all null hypotheses corresponding to $p_1, p_2, \ldots, p_k$.

Pro: much more accurate

Con: not super easy to wrap your head around.

Multiple Testing Problem

Do we ALWAYS need to correct for multiple testing?

This would be my suggestion to you:

- If you are seeking to conclude that there is an association between X and Y, use your favorite method of correeection (mine is the Benjamini-Hochberg)
- If you are looking for candidates, and plan to follow up, you may consider to not correct

(For simplicity, assume we're testing $H_0 : \mu = 17$ vs. $H_A : \mu \neq 17$ at a $\alpha = 0.05$ level of significance.)
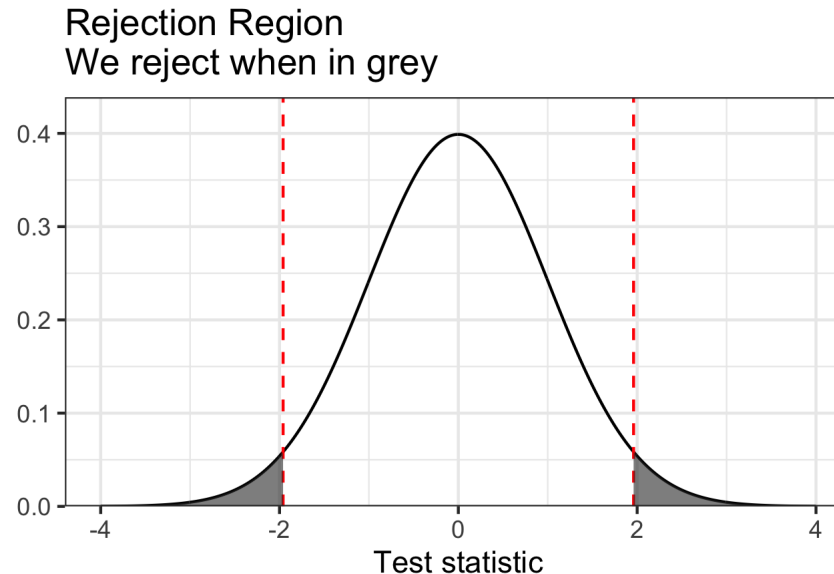
Get a sample

Calculate $t_{obs} = \frac{\bar{x} - 17}{\text{SD}(\bar{x})}$

Find $P\left(T > |t_{obs}| \mid H_0 \text{ is true}\right)$. This is our p-value. Compare to significance level. If smaller, reject. If larger, do not reject.

Can also ask: what value of T will give us p-value smaller than α?
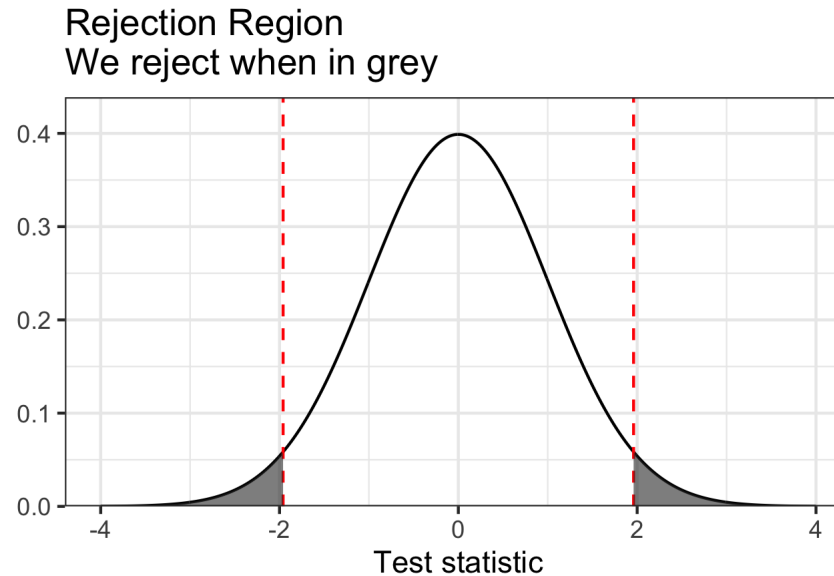

Rejection Region
We reject when in grey

All values in the grey: if we are outside the red dotted lines, we are "far away" from the null. Hence, we reject.

But that also means, we do NOT reject when in between red dotted lines.

Rejection Region
We reject when in grey

Confidence Intervals

Statistical Hypothesis Testing: is this one value possibly the truth?

Confidence Intervals: what values do we most believe to possibly be the truth?

Imagine you test a range of different hypothesis.

Confidence interval: the values that are NOT rejected.

So where the hypothesis test is rather pessimistic and inefficient, the confidence interval provides a whole lot more information!
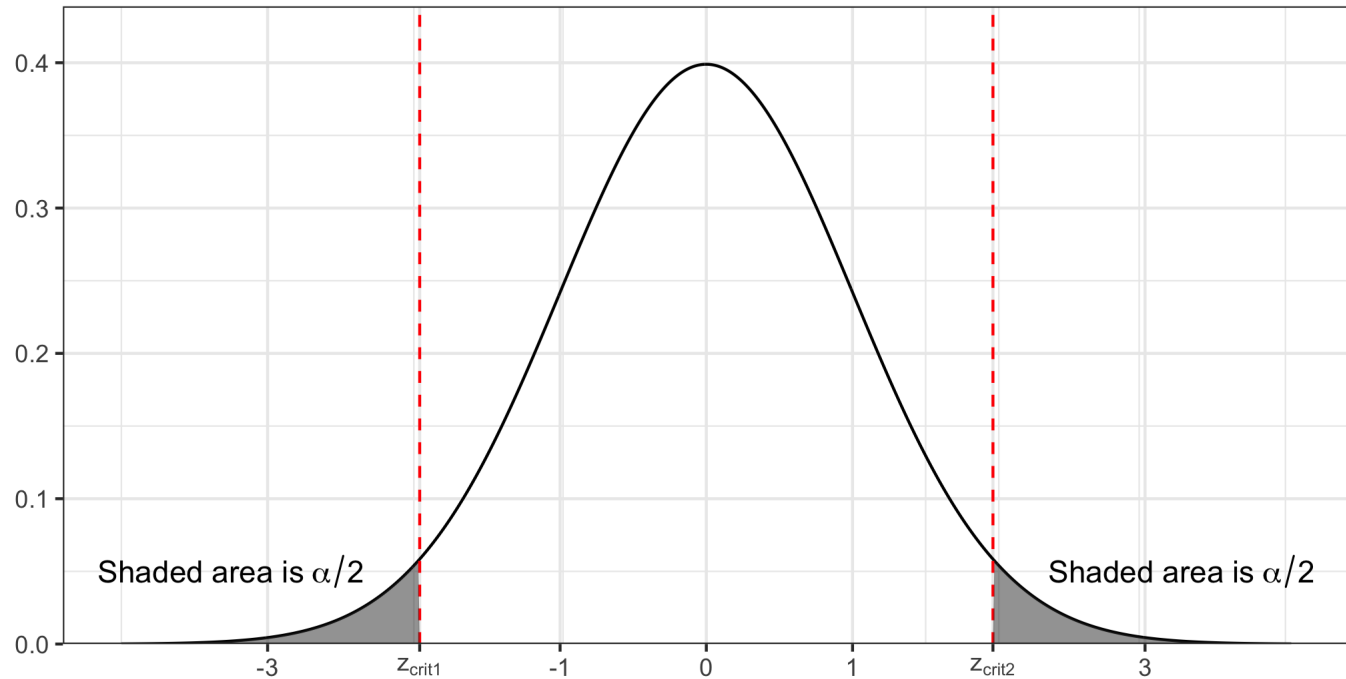
The question then is: why would you ever perform a test, and report a p-value, but not a confidence interval? I have no idea..............

The picture to find the critical values:

critical values = cut off $\alpha/2$ in the tails.

Coming up: the math

The math: we do not reject when $Z_{\text{crit}1} < Z_{\text{obs}} < Z_{\text{crit}2}$. Looking at the first inequality

$$z_{\text{crit},1} < z_{\text{obs}} \qquad\qquad \Longleftrightarrow$$

$$z_{\text{crit},1} < \frac{XX - \mu_0}{\text{SD}(XX)} \qquad\qquad \Longleftrightarrow$$

$$z_{\text{crit},1}\text{SD}(XX) < XX - \mu_0 \qquad\qquad \Longleftrightarrow$$

$$\mu_0 + z_{\text{crit},1} \cdot \text{SD}(XX) < XX \qquad\qquad \Longleftrightarrow$$

$$\mu_0 < XX - z_{\text{crit},1} \cdot \text{SD}(XX)$$

Looking at the second inequality:

$$z_{\mathrm{obs}} < z_{\mathrm{crit},2} \qquad\qquad \Longleftrightarrow$$

$$\frac{XX - \mu_0}{\mathrm{SD}(XX)} < z_{\mathrm{crit},2} \qquad\qquad \Longleftrightarrow$$

$$XX - \mu_0 < z_{\mathrm{crit},2} \cdot \mathrm{SD}(XX) \qquad\qquad \Longleftrightarrow$$

$$XX < \mu_0 + z_{\mathrm{crit},2} \cdot \mathrm{SD}(XX) \qquad\qquad \Longleftrightarrow$$

$$XX - z_{\mathrm{crit},2} \cdot \mathrm{SD}(XX) < \mu_0$$

So a $(1 - \alpha) \cdot 100\%$ Confidence Interval for μ is given by

$$[XX - z_{\text{crit},2} \cdot \text{SD}(XX), XX - z_{\text{crit},1} \cdot \text{SD}(XX)]$$

Whenever the distribution we're looking at is symmetrical (for example, the normal distribution or a t-distribution), then $z_{\text{crit},1} = -z_{\text{crit},2}$.

So, we can rewrite this as $XX \pm z_{\text{crit}} \cdot \text{SD}(XX)$.

What exactly is z_{crit}? The value that cuts off $\alpha/2$ of the (often normal) distribution. If $\alpha = 0.05$, then $z_{\text{crit}} = 1.96$.

Why do we call this $(1 - \alpha) \cdot 100\%$ (if $\alpha = 0.05$, this would be 95%) Confidence Interval?

$$P(XX - 1.96 \cdot \text{SD}(XX) < \mu < XX + 1.96 \cdot \text{SD}(XX))$$

$$= P(\mu < XX + 1.96 \cdot \text{SD}(XX)) - P(\mu < XX - 1.96 \cdot \text{SD}(XX))$$

$$= P(\mu - 1.96 \cdot \text{SD}(XX) < XX) - P(\mu - 1.96 \cdot \text{SD}(XX) < XX)$$

$$= P(-1.96 \cdot \text{SD}(XX) < XX - \mu) - P(XX - \mu < 1.96 \cdot \text{SD}(XX))$$

$$= P(-1.96 < \frac{XX - \mu}{\text{SD}(XX)}) - P(\frac{XX - \mu}{\text{SD}(XX)} < 1.96)$$

$$= P(-1.96 < Z) - P(Z < 1.96)$$

$$= [1 - P(Z < -1.96)] - P(Z < 1.96)$$

$$= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.$$

Confidence Intervals

Confidence Intervals

So what is this good for? How do we interpret a confidence interval?

CANNOT SAY: "The true value is in the confidence interval with probability 95%."

DO SAY: "We're 95% confident that the true value is in the confidence interval."

We'll see a bunch of examples on Tuesday.