6. Confidence Intervals

Dave Goldsman

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology

1/15/20

Georgia Tech

Outline

Lesson 6.1 — Introduction to Confidence Intervals

Next Few Lessons:

- This Introduction to Confidence Intervals.
- CI for the Mean of a Normal with *Known* Variance (easiest case).
- CI for the Mean Difference of Two Normals with *Known* Variances.

Idea: Instead of estimating a parameter by a **point estimator** alone, give a (random) **interval** that contains the parameter with a certain probability.

Example: $\bar{X}$ is a point estimator for the parameter μ.
A 95% **confidence interval** for μ might look like

$$\mu \in \left[\bar{X} - z_{0.025}\sqrt{\sigma^2/n},\ \ \bar{X} + z_{0.025}\sqrt{\sigma^2/n}\right],$$

where $z_{0.025}$ is the $\text{Nor}(0,1)$'s 0.975 quantile.
This means that μ is in the interval with probability 0.95.

Definition: A $100(1-\alpha)\%$ **confidence interval** for an unknown parameter θ is given by two random variables L and U satisfying

$$P(L \le \theta \le U) \;=\; 1-\alpha.$$

L and U are the **lower** and **upper** confidence limits, and $1-\alpha$ is the **confidence coefficient**, specified in advance.

There is a $1-\alpha$ chance that θ actually lies between L and U.

Example: We're 95% sure that President Smith's popularity is $56\% \pm 3\%$.

Since $L \le \theta \le U$, we call $[L, U]$ a **two-sided** CI for θ.

If L is such that $P(L \le \theta) = 1-\alpha$, then $[L, \infty)$ is a $100(1-\alpha)\%$ **one-sided lower** CI for θ.

Similarly, if U is such that $P(\theta \le U) = 1-\alpha$, then $(-\infty, U]$ is a $100(1-\alpha)\%$ **one-sided upper** CI for θ.

Example: Here are some results from 10 independent samples, each consisting of 100 different observations. From each sample, we use the 100 observations to re-calculate L and U.

Is the unknown θ in $[L, U]$?

Sample #	L	U	θ	CI covers θ?
1	1.86	2.23	2	Yes
2	1.90	2.31	2	Yes
3	3.21	3.86	2	No
4	1.75	2.10	2	Yes
5	1.72	2.03	2	Yes
$\vdots$	$\vdots$	$\vdots$	$\vdots$	
10	1.62	1.98	2	No

As the number of samples gets large, the proportion of CIs that cover the unknown θ approaches $1-\alpha$.

Sometimes CIs miss θ too low, i.e., $\theta > U$.

Sometimes CIs miss θ too high, i.e., $\theta < L$.

But $1-\alpha$ of the time, they're just right — Three Little Bears.

There are lots of different kinds of confidence intervals. We'll look at CIs for the mean and variance of normal distributions as well as CIs for the success probability of Bernoulli distributions. We'll also extend those results to compare competing normal distributions as well as competing Bernoulli distributions.

Outline

Lesson 6.2 — Normal Mean (variance known)

Set-up: Sample from a normal distribution with unknown mean μ and *known* variance σ^2.

Goal: Obtain a confidence interval for μ.

Remark: This is an unrealistic case, since if we didn't know μ in real life, then we wouldn't know σ^2 either. But it's a good place to start the discussion.

Details: Suppose $X_1, \dots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, where σ^2 is *known*.

Use $\bar{X} = \sum_{i=1}^n X_i/n$ as our point estimator. Recall

$$\bar{X} \sim \text{Nor}(\mu, \sigma^2/n) \Rightarrow Z \equiv \frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \text{Nor}(0,1).$$

The quantity Z is called a **pivot**. It's a "starting point" for us.

The definition of Z implies that

$$
\begin{aligned}
1-\alpha &= P\Big(-z_{\alpha/2} \le Z \le z_{\alpha/2}\Big) \\
&= P\Big(-z_{\alpha/2} \le \frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \le z_{\alpha/2}\Big) \\
&= P\Big(-z_{\alpha/2}\sqrt{\sigma^2/n} \le \bar{X}-\mu \le z_{\alpha/2}\sqrt{\sigma^2/n}\Big) \\
&= P\Big(\underbrace{\bar{X}-z_{\alpha/2}\sqrt{\sigma^2/n}}_{L} \le \mu \le \underbrace{\bar{X}+z_{\alpha/2}\sqrt{\sigma^2/n}}_{U}\Big) \\
&= P(L \le \mu \le U).
\end{aligned}
$$

Thus, the $100(1-\alpha)\%$ two-sided CI for μ is

$$
\boxed{\bar{X}-z_{\alpha/2}\sqrt{\sigma^2/n} \;\le\; \mu \;\le\; \bar{X}+z_{\alpha/2}\sqrt{\sigma^2/n}.}
$$

Remarks:

Notice how we used the pivot to "isolate" μ all by itself to the middle of the inequalities.

After you observe $X_1, \ldots, X_n$, you calculate L and U. Nothing is unknown, since L and U don't involve μ.

Sometimes we'll write the CI as $\mu \in \bar{X} \pm H$, where the **half-width** is

$$H \equiv z_{\alpha/2}\sqrt{\sigma^2/n}.$$

Example: Suppose we take $n = 25$ iid observations from a $\text{Nor}(\mu, \sigma^2)$ distribution, where we somehow *know* that $\sigma = 30$. Further suppose that $\bar{X}$ turns out to be 278. Let's find a $100(1-\alpha)\% = 95\%$ CI for μ.

$$\begin{aligned} \mu &\in \bar{X} \pm z_{\alpha/2}\sqrt{\sigma^2/n} \\ &= 278 \pm z_{0.025}(30/5) \\ &= 278 \pm 11.76 \quad (z_{0.025} = 1.96). \end{aligned}$$

So a 95% CI for μ is $266.24 \le \mu \le 289.76$. $\Box$

Sample-Size Calculation

If we had taken more observations, then the CI would have gotten shorter, since $H = z_{\alpha/2}\sqrt{\sigma^2/n}$.

In fact, how many observations should be taken to make the half-length (or "error") $\leq \epsilon$?

$$z_{\alpha/2}\sqrt{\sigma^2/n} \;\leq\; \epsilon \quad \text{iff} \quad n \;\geq\; (\sigma z_{\alpha/2}/\epsilon)^2.$$

Thus, the sample size goes up as…

- The variance σ^2 increases (more uncertainty).
- The required half-width decreases (becomes more stringent).
- The value of α decreases (more confidence).

Example: Suppose, in the previous example, that we want the half-length to be ≤ 10, i.e., $\mu \in \bar{X} \pm 10$. What should n be?

$$n \;\geq\; (\sigma z_{\alpha/2}/\epsilon)^2 \;=\; \big((30)(1.96)/10\big)^2 \;=\; 34.57.$$

Just to make n an integer, round up to $n = 35$. $\square$

Remark: We can similarly obtain one-sided CIs for μ (if we're just interested in one bound)...

$100(1-\alpha)\%$ upper CI: $\mu \;\leq\; \bar{X} + z_\alpha\sqrt{\sigma^2/n}$.

$100(1-\alpha)\%$ lower CI: $\mu \;\geq\; \bar{X} - z_\alpha\sqrt{\sigma^2/n}$.

Note that we use the $1-\alpha$ quantile (not $1-\alpha/2$).

Outline

Lesson 6.3 — Difference of Two Normal Means (variances known)

Idea: Compare the means of two competing alternatives or processes by getting a confidence interval for their difference.

Example: Do Georgia Tech students have higher average IQ's than Univ. of Georgia students? Answer: Yes.

We'll again assume that the variances of the two competitors are somehow known. We'll do the more-realistic unknown variance case in the next lesson.

Set-up: Suppose we have samples of sizes n and m from the two competing populations.

$$\begin{aligned} X_1, X_2, \ldots, X_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{(population 1)} \\ Y_1, Y_2, \ldots, Y_m &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2) \quad \text{(population 2)}, \end{aligned}$$

where the means μ_x and μ_y are *unknown*, while σ_x^2 and σ_y^2 are somehow *known*.

Also assume that the X_i's are *independent* of the Y_i's.

Let's find a CI for the difference in means, $\mu_x - \mu_y$.

Define the sample means from populations 1 and 2,

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

Obviously,

$$\bar{X} \sim \text{Nor}(\mu_x, \sigma_x^2/n) \quad \text{and} \quad \bar{Y} \sim \text{Nor}(\mu_y, \sigma_y^2/m),$$

so that

$$\bar{X} - \bar{Y} \sim \text{Nor}\left(\mu_x - \mu_y,\ \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}\right).$$

This implies that

$$Z \equiv \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}} \sim \text{Nor}(0,1),$$

so that

$$P(-z_{\alpha/2} \le Z \le z_{\alpha/2}) = 1 - \alpha.$$

Using the same manipulations as in the single-population case, we get a two-sided $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$:

$$\boxed{\mu_x - \mu_y \in \bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}.}$$

Similarly,

One-sided upper CI:

$$\mu_x - \mu_y \ \leq \ \bar{X} - \bar{Y} + z_\alpha \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}.$$

One-sided lower CI:

$$\mu_x - \mu_y \ \geq \ \bar{X} - \bar{Y} - z_\alpha \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}}.$$

Example: A traveling professor gives the same test to Georgia Tech (X) and University of Georgia (Y) students. She assumes that the test scores are normally distributed with known standard deviations of $\sigma_x = 20$ points and $\sigma_y = 12$ points, respectively. She takes random samples of 40 GT scores and 24 UGA tests and observes sample means of $\bar{X} = 95$ points and $\bar{Y} = 60$, respectively. Let's find the 90% two-sided CI for $\mu_x - \mu_y$.

$$\begin{aligned} \mu_x - \mu_y &\in \bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \\ &= 35 \pm 1.645\sqrt{\frac{400}{40} + \frac{144}{24}} \\ &= 35 \pm 6.58, \end{aligned}$$

implying that $28.42 \le \mu_x - \mu_y \le 41.58$. In other words, we're 90% sure that $\mu_x - \mu_y$ lies in this interval. $\Box$

Remark: Let's assume that the sample sizes are equal, i.e., $n = m$. To obtain a half-length $\leq \epsilon$, we require

$$n \ \geq \ \frac{z_{\alpha/2}^2(\sigma_x^2 + \sigma_y^2)}{\epsilon^2}.$$

In the next lesson we'll explore the more-realistic case in which the variance of the underlying observations is unknown. Now it's time for $t\ldots$.

Outline

Lesson 6.4 — Normal Mean (variance unknown)

Next Few Lessons:

- Confidence Interval for the Mean of a Normal Distribution with *Unknown* Variance.
- CIs for the Mean Difference of Two Normals with *Unknown* Variances when...
  - the variances are assumed to be *equal*.
  - the variances are *unequal*.
  - observations between the two normals are taken in special *pairs*.

At this point we'll look at the more-realistic case in which the variance of the underlying normal random variables is *unknown*.

This takes a little more work, but has many more applications.

Set-up: $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2)$, with σ^2 unknown.

Facts:

(a) $$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim \text{Nor}(0,1).$$

(b) $\bar{X}$ and S^2 are *independent*.

(c) $$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \frac{\sigma^2\chi^2(n-1)}{n-1}.$$

Georgia Tech

Sketch of Proof: We proved Fact (a) previously. The proof of Fact (b) uses what is known as Cochran's Theorem, which is a bit beyond our scope. In order to derive Fact (c), we first do some algebra to re-write $\sum_{i=1}^n (X_i-\mu)^2$,

$$\begin{aligned}
\sum_{i=1}^n (X_i-\mu)^2 &= \sum_{i=1}^n \left[(X_i-\bar{X})+(\bar{X}-\mu)\right]^2 \\
&= \sum_{i=1}^n (X_i-\bar{X})^2 + 2(\bar{X}-\mu)\sum_{i=1}^n (X_i-\bar{X}) + n(\bar{X}-\mu)^2 \\
&= \sum_{i=1}^n (X_i-\bar{X})^2 + n(\bar{X}-\mu)^2 \quad \Big(\textstyle\sum_i (X_i-\bar{X})=0\Big),
\end{aligned}$$

which can be rewritten as

$$\frac{1}{n-1}\sum_{i=1}^n (X_i-\mu)^2 \;=\; S^2 + \frac{n}{n-1}(\bar{X}-\mu)^2. \tag{1}$$

Now note that $X_i \sim \mathrm{Nor}(\mu, \sigma^2)$ and the fact that $[\mathrm{Nor}(0,1)]^2 \sim \chi^2(1)$ together imply $(X_i - \mu)^2 \sim \sigma^2\chi^2(1)$, for all i; so the additive property of independent χ^2's implies that

$$\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \mu)^2 \;\sim\; \frac{\sigma^2\chi^2(n)}{n-1}. \tag{2}$$

Similarly, Fact (a) implies that

$$\frac{n}{n-1}(\bar{X} - \mu)^2 \;\sim\; \frac{\sigma^2\chi^2(1)}{n-1}. \tag{3}$$

Since $\bar{X}$ and S^2 are independent (by Fact (b)), we can substitute the results from Equations (2) and (3) into (1) and then invoke χ^2 additivity to obtain

$$\frac{\sigma^2\chi^2(n)}{n-1} \;\sim\; S^2 + \frac{\sigma^2\chi^2(1)}{n-1} \quad\Rightarrow\quad S^2 \;\sim\; \frac{\sigma^2\chi^2(n-1)}{n-1}. \quad \Box$$

Using Facts (a)–(c), we have, by the definition of the t distribution,

$$\frac{\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}}}{\sqrt{S^2/\sigma^2}} \sim \frac{\mathrm{Nor}(0,1)}{\sqrt{\chi^2(n-1)/(n-1)}} \sim t(n-1).$$

In other words,

$$\frac{\bar{X}-\mu}{\sqrt{S^2/n}} \sim t(n-1).$$

Note that this expression doesn't contain the unknown σ^2. It's been replaced by S^2, which we can calculate.

This process is called **"standardizing and Studentizing."**

Georgia Tech

Now, by the **same** manipulations as in the known-variance case, we can get a two-sided $100(1-\alpha)\%$ CI for μ:

$$\boxed{\mu \ \in \ \bar{X} \pm t_{\alpha/2,n-1}\sqrt{S^2/n}.}$$

$100(1-\alpha)\%$ lower CI: $\mu \geq \bar{X} - t_{\alpha,n-1}\sqrt{S^2/n}$.

$100(1-\alpha)\%$ upper CI: $\mu \leq \bar{X} + t_{\alpha,n-1}\sqrt{S^2/n}$.

Remark: Here we use t-distribution quantiles (instead of normal quantiles as in the known-variance case from the previous lesson). The t quantile tends to be larger than the corresponding Nor(0,1) quantile, so these unknown-variance CIs tend to be a bit longer than the known-variance CIs. The longer CIs are the result of the fact that we lack precise info about the variance.

Example: Here are 20 residual flame times (in seconds) of treated specimens of children's nightwear. (Don't worry — children were not in the nightwear when the clothing was set on fire.)

9.85	9.93	9.75	9.77	9.67
9.87	9.67	9.94	9.85	9.75
9.83	9.92	9.74	9.99	9.88
9.95	9.95	9.93	9.92	9.89

Let's get a 95% CI for the mean residual flame time.

After a little algebra, we obtain

$$\bar{X} \;=\; 9.8525 \quad \text{and} \quad S \;=\; 0.0965.$$

Further, you can use the Excel function `T.INV(0.975,19)` to get $t_{\alpha/2,n-1} = t_{0.025,19} = 2.093$.

Then the half-length of the CI is

$$H \;=\; t_{\alpha/2,n-1}\sqrt{S^2/n} \;=\; \frac{(2.093)(0.0965)}{\sqrt{20}} \;=\; 0.0451.$$

Thus, the CI is $\mu \in \bar{X} \pm H$, or $9.8074 \le \mu \le 9.8976$. $\Box$

R Code:

```
x <- data.frame(x=c(
9.85, 9.93, 9.75, 9.77, 9.67,
9.87, 9.67, 9.94, 9.85, 9.75,
9.83, 9.92, 9.74, 9.99, 9.88,
9.95, 9.95, 9.93, 9.92, 9.89))

print(confint(lm(x~1,x)))
               2.5 \%   97.5 \%
(Intercept) 9.807357 9.897643
```

Outline

Lesson 6.5 — Difference of Two Normal Means (unknown *equal* variances)

Compare the means from two competing normal populations.

Suppose we have *independent* samples of sizes n and m from the two.

$$\begin{aligned} X_1, X_2, \ldots, X_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2) \quad \text{(population 1).} \\ Y_1, Y_2, \ldots, Y_m &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2) \quad \text{(population 2).} \end{aligned}$$

We assume that the means μ_x and μ_y are *unknown*, and the variances σ_x^2 and σ_y^2 are also *unknown*.

Big Assumption: Suppose for now that $\sigma_x^2 = \sigma_y^2 = \sigma^2$, i.e., that the two variances are *unknown* but *equal*.

Let's find a CI for the difference in means, $\mu_x - \mu_y$.

To get things going, calculate sample means.

$$\bar{X} \equiv \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{and} \quad \bar{Y} \equiv \frac{1}{m}\sum_{i=1}^{m} Y_i.$$

Now, sample variances.

$$S_x^2 \equiv \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \quad \text{and} \quad S_y^2 \equiv \frac{\sum_{i=1}^{m}(Y_i - \bar{Y})^2}{m-1}.$$

Both S_x^2 and S_y^2 are estimators for σ^2 (the common variance).

A better estimator is the **pooled** estimator for σ^2, which uses the info from both S_x^2 and S_y^2.

$$S_p^2 \equiv \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2}.$$

Theorem:

$$S_p^2 \sim \frac{\sigma^2 \chi^2(n+m-2)}{n+m-2},$$

Proof: Follows from the fact that independent χ^2 RVs add up. $\Box$

After some of the usual algebra, it turns out that

$$\frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{S_p \sqrt{\frac{1}{n} + \frac{1}{m}}} \sim t(n+m-2).$$

So, after more of the usual algebra, we get a two-sided $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$:

$$\boxed{\mu_x - \mu_y \;\in\; \bar{X} - \bar{Y} \pm t_{\alpha/2,n+m-2}\, S_p \sqrt{\frac{1}{n} + \frac{1}{m}},}$$

where $t_{\alpha/2,n+m-2}$ is the appropriate t distribution quantile.

One-Sided Lower CI:

$$\mu_x - \mu_y \;\geq\; \bar{X} - \bar{Y} - t_{\alpha,n+m-2}\, S_p \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

One-Sided Upper CI:

$$\mu_x - \mu_y \;\leq\; \bar{X} - \bar{Y} + t_{\alpha,n+m-2}\, S_p \sqrt{\frac{1}{n} + \frac{1}{m}}.$$

Example: IQ's of students at Georgia Tech and the "Univ." of Georgia.

Georgia Tech students: $X_1, \ldots, X_{25} \stackrel{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma^2)$.

Univ. of Georgia students: $Y_1, \ldots, Y_{36} \stackrel{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma^2)$.

Note: We assume common σ^2.

Suppose it turns out that

$$\bar{X} = 120 \quad \text{and} \quad \bar{Y} = 80.$$

$$S_x^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = 100.$$

$$S_y^2 = \frac{\sum_{i=1}^m (Y_i - \bar{Y})^2}{m-1} = 95.$$

The two sample variances are pretty close, so we'll go ahead and feel good about our common σ^2 assumption; and we'll use the pooled estimator,

$$S_p^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{n+m-2} = \frac{(24)(100) + (35)(95)}{59} = 97.03.$$

Thus, a two-sided 95% CI for $\mu_x - \mu_y$ is

$$\begin{aligned} \mu_x - \mu_y &\in \bar{X} - \bar{Y} \pm t_{\alpha/2,n+m-2}\, S_p \sqrt{\frac{1}{n} + \frac{1}{m}} \\ &= 120 - 80 \pm 2.00\sqrt{97.03}\sqrt{0.0678} \\ &= 40 \pm 5.13. \end{aligned}$$

So the 95% CI is $34.87 \le \mu_x - \mu_y \le 45.13$.

Note that the above CI doesn't contain 0 (not even close). But it's obvious anyway that GT students are smarter than UGA students! Duh. $\square$

Outline

Lesson 6.6 — Difference of Two Normal Means (variances unknown)

Again find a CI for the difference in means, $\mu_x - \mu_y$, from two normal populations. But now the normals have *unknown* and possibly *unequal* variances, σ_x^2 and σ_y^2.

Again suppose we have samples of sizes n and m,

$$\begin{array}{rcll} X_1, X_2, \ldots, X_n & \stackrel{\text{iid}}{\sim} & \text{Nor}(\mu_x, \sigma_x^2) & \text{(population 1)} \\ Y_1, Y_2, \ldots, Y_m & \stackrel{\text{iid}}{\sim} & \text{Nor}(\mu_y, \sigma_y^2) & \text{(population 2),} \end{array}$$

where we assume that the X_i's are *independent* of the Y_j's.

As before, start by calculating sample means and variances, $\bar{X}$, $\bar{Y}$, S_x^2, S_y^2.

Since the variances are possibly unequal, we can't use the pooled estimator S_p^2. Instead, use **Welch's approximation method**.

$$t^\star \equiv \frac{\bar{X} - \bar{Y} - (\mu_x - \mu_y)}{\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}} \approx t(\nu),$$

where the *approximate* degrees of freedom is given by

$$\nu \equiv \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}}.$$

After the usual algebra, we obtain an approximate two-sided $100(1-\alpha)\%$ CI for $\mu_x - \mu_y$ (which has very wide application in practice):

$$\boxed{\mu_x - \mu_y \in \bar{X} - \bar{Y} \pm t_{\alpha/2,\nu} \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}.}$$

Georgia Tech

Example: Two normal populations. Let's get a 95% CI for $\mu_x - \mu_y$. Suppose it turns out that

$$\begin{array}{ccc} n = 25 & \bar{X} = 100 & S_x^2 = 400 \\ m = 16 & \bar{Y} = 80 & S_y^2 = 100. \end{array}$$

You can tell from S_x^2 and S_y^2 that there's no way that the two variances are equal. So we'll have to use the approximation method.

The approximate degrees of freedom is

$$\nu \equiv \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{m}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}} = \frac{\left(\frac{400}{25} + \frac{100}{16}\right)^2}{\frac{(400/25)^2}{24} + \frac{(100/16)^2}{15}} = 37.30 \doteq 37,$$

where we round df *down* to be *conservative* (i.e., slightly longer CIs).

Since the confidence coefficient is 0.95, we have $t_{\alpha/2,\nu} = t_{0.025,37} = 2.026$.

This gives us the following CI for $\mu_x - \mu_y$.

$$\begin{aligned}\mu_x - \mu_y &\in \bar{X} - \bar{Y} \pm t_{\alpha/2,\nu}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}} \\ &= 20 \pm 2.026\sqrt{22.25} \\ &= 20 \pm 9.56. \quad \Box\end{aligned}$$

Outline

Lesson 6.7 — Difference of Paired Normal Means (variances unknown)

Again consider two competing normal populations. Suppose we collect observations from the two populations in *pairs*.

The RVs *between different pairs* are *independent*. The two observations *within the same pair* may *not* be independent.

$$\text{independent}\left\{\begin{array}{ll} \text{Pair } 1: & (X_1, Y_1) \\ \text{Pair } 2: & (X_2, Y_2) \\ \vdots & \vdots \\ \text{Pair } n: & \underbrace{(X_n, Y_n)}_{\text{not indep}} \end{array}\right.$$

Pair i is indep of pair j, i.e., the pair (X_i, Y_i) is indep of (X_j, Y_j). But within pair i, it may be that X_i and Y_i are not independent.

Example: One twin takes a new drug, the other takes a placebo.

Idea: By setting up such experiments, we hope to be able to capture the difference between the two normal populations more precisely, since we're using the pairs to eliminate extraneous noise.

Here's the set-up. Take n pairs of observations:

$$\begin{aligned} X_1, X_2, \ldots, X_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2). \\ Y_1, Y_2, \ldots, Y_n &\overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2). \end{aligned}$$

(Additional technical assumption: All X_i's and Y_j's are jointly normal.)

We assume that the means μ_x and μ_y are *unknown*, and the variances σ_x^2 and σ_y^2 are also *unknown* and possibly *unequal*.

Further, pair i is independent of pair j (between pairs), but X_i may not be independent of Y_i (within a pair).

Goal: Find a CI for the difference in means, $\mu_x - \mu_y$.

Define the pair-wise differences,

$$D_i \;\equiv\; X_i - Y_i, \quad i = 1, 2, \ldots, n.$$

Note that

$$D_1, D_2, \ldots, D_n \;\stackrel{\text{iid}}{\sim}\; \text{Nor}(\mu_d, \sigma_d^2),$$

where $\mu_d \equiv \mu_x - \mu_y$ (which is what we want the CI for), and

$$\sigma_d^2 \;\equiv\; \sigma_x^2 + \sigma_y^2 - 2\,\text{Cov}(X_i, Y_i).$$

Now the problem reduces to the old $\text{Nor}(\mu, \sigma^2)$ case with unknown μ and σ^2.

So let's calculate the sample mean and variance as before.

$$\bar{D} \equiv \frac{1}{n}\sum_{i=1}^{n} D_i \sim \mathrm{Nor}(\mu_d, \sigma_d^2/n).$$

$$S_d^2 \equiv \frac{1}{n-1}\sum_{i=1}^{n}(D_i - \bar{D})^2 \sim \frac{\sigma_d^2\,\chi^2(n-1)}{n-1}.$$

Just like before, we get

$$\frac{\bar{D} - \mu_d}{\sqrt{S_d^2/n}} \sim t(n-1).$$

After the usual algebra on the pivot (with the t distribution instead of the normal), we obtain a two-sided $100(1-\alpha)\%$ CI,

$$\boxed{\mu_d \;=\; \mu_x - \mu_y \;\in\; \bar{D} \pm t_{\alpha/2,n-1}\,\sqrt{S_d^2/n}.}$$

One-sided lower: $\mu_d \geq \bar{D} - t_{\alpha,n-1}\,\sqrt{S_d^2/n}$.

One-sided upper: $\mu_d \leq \bar{D} + t_{\alpha,n-1}\,\sqrt{S_d^2/n}$.

Example: Times for people to parallel park two cars.

Person	Park Honda	Park Cadillac	Difference
1	10	20	-10
2	25	40	-15
3	5	5	0
4	20	35	-15
5	15	20	-5

Clearly, the people are independent, but the times for the same individual to park the two cars may not be independent.

Let's assume that all times are normal. We want a 90% two-sided CI for $\mu_d = \mu_h - \mu_c$.

We see that $n = 5$, $\bar{D} = -9$, and $S_d^2 = 42.5$.

Thus, the CI is

$$\begin{aligned} \mu_d &\in \bar{D} \pm t_{0.05,4} \sqrt{S_d^2/n} \\ &= -9 \pm 2.13 \sqrt{42.5/5} \\ &= -9 \pm 6.21. \quad \Box \end{aligned}$$

So why didn't we just use the "usual" CI for the difference of two means? Namely,

$$\mu_x - \mu_y \ \in \ \bar{X} - \bar{Y} \pm t_{\alpha/2,\nu} \sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{m}}.$$

Main reason: This CI requires the X's to be independent of the Y's. (Recall that the paired-t method allows X_i and Y_i to be dependent.)

Good thing about using the "usual" method: The approximate df ν from the "usual" method would probably be larger than the df $n-1$ from the paired-t method. This would make the CI smaller.

Bad thing: You'd introduce much more noise into the system by using the "usual" method. This could make the CI much larger.

Example: Back to the car example (now use "usual" approximate method).

A guy parks Honda X_i	Same guy parks Caddy	Different guy parks Caddy Y_i
10	20	30
25	40	15
5	5	40
20	35	10
15	20	25

Just concentrate on the X_i and Y_i columns. (The middle column is from the last example for comparison purposes.)

The X_i's and Y_i's have the same sample averages as before, but now there's more natural variation, since we're using 10 different people.

Now all of the X_i's are independent of all of the Y_i's.

$$\bar{X} = 15, \quad \bar{Y} = 24, \quad S_x^2 = 62.5, \quad S_y^2 = 142.5.$$

Then we have

$$\nu = \frac{\left(\frac{S_x^2}{n} + \frac{S_y^2}{n}\right)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/n)^2}{n-1}} = \frac{4\big(62.5 + 142.5\big)^2}{(62.5)^2 + (142.5)^2} = 6.94 \doteq 7.$$

This gives us the following 90% CI for $\mu_x - \mu_y$.

$$\mu_x - \mu_y \in \bar{X} - \bar{Y} \pm t_{0.05,7}\sqrt{\frac{S_x^2}{n} + \frac{S_y^2}{n}} = -9 \pm 1.895\sqrt{41} = -9 \pm 12.13.$$

This CI is *wider* than the paired-t version, even though we have more df here.

Moral: Use paired-t when you can. $\square$

Georgia Tech

Outline

Lesson 6.8 — Normal Variance

Next Few Lessons: We'll now look at a potpourri of different types of CIs.

- Normal variance.
- Ratio of normal variances from two populations.
- Bernoulli proportion.
- Difference of Bernoulli proportions from two populations.

First, CIs for normal variance....

To address how much variability we can expect from some system, we'll now get a CI for the variance σ^2 of a normal distribution (instead of the mean).

Usual set-up:

$$X_1, X_2, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Nor}(\mu, \sigma^2).$$

Recall that the distribution of the sample variance is

$$S^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \;\sim\; \frac{\sigma^2\chi^2(n-1)}{n-1}.$$

Since

$$\frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1),$$

we use χ^2 quantiles and some algebra to get

$$\begin{aligned}
1-\alpha &= P\left(\chi^2_{1-\frac{\alpha}{2},n-1} \le \frac{(n-1)S^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2},n-1}\right) \\
&= P\left(\frac{1}{\chi^2_{1-\frac{\alpha}{2},n-1}} \ge \frac{\sigma^2}{(n-1)S^2} \ge \frac{1}{\chi^2_{\frac{\alpha}{2},n-1}}\right) \\
&= P\left(\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}} \le \sigma^2 \le \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right).
\end{aligned}$$

So a $100(1-\alpha)\%$ CI for σ^2 is

$$\boxed{\sigma^2 \ \in \ \left[\frac{(n-1)S^2}{\chi^2_{\frac{\alpha}{2},n-1}}, \ \ \frac{(n-1)S^2}{\chi^2_{1-\frac{\alpha}{2},n-1}}\right].}$$

Remark: The CI for σ^2 is directly proportional to the sample variance S^2.

Remark: This CI contains no reference to the unknown μ!

Meanwhile, a $100(1-\alpha)\%$ lower CI for σ^2 is:

$$\frac{(n-1)S^2}{\chi^2_{\alpha,n-1}} \ \le \ \sigma^2.$$

$100(1-\alpha)\%$ upper CI for σ^2:

$$\sigma^2 \ \le \ \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}}.$$

Example: Suppose 25 people take an IQ test and that their scores are normally distributed.

If $S^2 = 100$, find a 95% upper CI for the variance σ^2.

Looking up the χ^2 quantile, we get

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{1-\alpha,n-1}} = \frac{(24)(100)}{\chi^2_{0.95,24}} = \frac{2400}{13.85} = 173.3. \quad \Box$$

Outline

Lesson 6.9 — Ratio of Variances of Two Normals

Which of two normal distributions is more variable?

Set-up:

$$X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Nor}(\mu_x, \sigma_x^2).$$
$$Y_1, Y_2, \ldots, Y_m \overset{\text{iid}}{\sim} \text{Nor}(\mu_y, \sigma_y^2).$$

All X's and Y's are independent with unknown means and variances.

We'll get a confidence interval for the *ratio* σ_x^2/σ_y^2.

Recall the distributions of the two sample variances:

$$S_x^2 \;=\; \frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2 \;\sim\; \frac{\sigma_x^2\,\chi^2(n-1)}{n-1}, \quad \text{and}$$

$$S_y^2 \;=\; \frac{1}{m-1}\sum_{i=1}^{m}(Y_i-\bar{Y})^2 \;\sim\; \frac{\sigma_y^2\,\chi^2(m-1)}{m-1}.$$

Thus,

$$\frac{S_y^2/\sigma_y^2}{S_x^2/\sigma_x^2} \;\sim\; \frac{\chi^2(m-1)/(m-1)}{\chi^2(n-1)/(n-1)} \;\sim\; F(m-1,n-1).$$

Using F quantiles and some algebra, we get

$$\begin{aligned} 1-\alpha &= P\left(F_{1-\frac{\alpha}{2},m-1,n-1} \le \frac{S_y^2/\sigma_y^2}{S_x^2/\sigma_x^2} \le F_{\frac{\alpha}{2},m-1,n-1}\right) \\ &= P\left(\frac{S_x^2}{S_y^2}\,\frac{1}{F_{\frac{\alpha}{2},n-1,m-1}} \le \frac{\sigma_x^2}{\sigma_y^2} \le \frac{S_x^2}{S_y^2}\,F_{\frac{\alpha}{2},m-1,n-1}\right). \end{aligned}$$

So a $100(1-\alpha)\%$ CI for σ_x^2/σ_y^2 is

$$\boxed{\frac{\sigma_x^2}{\sigma_y^2} \in \left[\frac{S_x^2}{S_y^2}\,\frac{1}{F_{\frac{\alpha}{2},n-1,m-1}},\ \ \frac{S_x^2}{S_y^2}\,F_{\frac{\alpha}{2},m-1,n-1}\right].}$$

Remark: The CI for σ_x^2/σ_y^2 is proportional to the ratio of the sample variances, S_x^2/S_y^2, and contains no reference to μ_x or μ_y.

Georgia Tech

Meanwhile, a $100(1-\alpha)\%$ lower CI for σ_x^2/σ_y^2 is:

$$\frac{S_x^2}{S_y^2}\,\frac{1}{F_{\alpha,n-1,m-1}} \;\le\; \frac{\sigma_x^2}{\sigma_y^2}.$$

$100(1-\alpha)\%$ upper CI for σ_x^2/σ_y^2:

$$\frac{\sigma_x^2}{\sigma_y^2} \;\le\; \frac{S_x^2}{S_y^2}\,F_{\alpha,m-1,n-1}.$$

Remark: If you want CIs for σ_y^2/σ_x^2, just flip all of the X's and Y's in the various CIs discussed above.

Example: Suppose 25 people take IQ test A, and 16 people take IQ test B. Assume all scores are normal and independent.

If $S_A^2 = 100$ and $S_B^2 = 70$, find a 95% upper CI for the ratio σ_A^2/σ_B^2.

Looking up the $F_{\alpha,n_B-1,n_A-1} = F_{0.05,15,24}$ quantile, we get

$$\frac{\sigma_A^2}{\sigma_B^2} \le \frac{S_A^2}{S_B^2} F_{\alpha,n_B-1,n_A-1} = \frac{100}{70}(2.11) = 3.01. \quad \square$$

Outline

Lesson 6.10 — Bernoulli Proportion

Suppose that $X_1, X_2, \ldots, X_n \overset{\text{iid}}{\sim} \text{Bern}(p)$.

What probability of "success" can we expect from this distribution?

We'll get a CI for the proportion p of successes.

Since $\sum_{i=1}^{n} X_i \sim \text{Bin}(n, p)$, we know that

$$\bar{X} \sim \frac{\text{Bin}(n,p)}{n}.$$

Let's assume that n is "large," so that we'll be able to use the Central Limit Theorem (and don't worry about the continuity correction).

(If n isn't large, then we'll have to use nasty Binomial tables, which I don't want to deal with here!)

Note that

$$\begin{aligned} \mathrm{E}[\bar{X}] &= \mathrm{E}[X_i] = p, \quad \text{and} \\ \mathrm{Var}(\bar{X}) &= \mathrm{Var}(X_i)/n = pq/n. \end{aligned}$$

Then for large n, the Central Limit Theorem implies

$$\frac{\bar{X} - \mathrm{E}[\bar{X}]}{\sqrt{\mathrm{Var}(\bar{X})}} = \frac{\bar{X} - p}{\sqrt{pq/n}} \approx \mathrm{Nor}(0,1).$$

Now let's do something crazy and estimate pq by its maximum likelihood estimator, $\bar{X}(1-\bar{X})$. This gives

$$\frac{\bar{X} - p}{\sqrt{\bar{X}(1-\bar{X})/n}} \approx \mathrm{Nor}(0,1).$$

Then the "usual" algebra implies

$$
\begin{aligned}
1-\alpha &\doteq P\left(-z_{\alpha/2} \le \frac{\bar{X}-p}{\sqrt{\bar{X}(1-\bar{X})/n}} \le z_{\alpha/2}\right) \\
&= P\left(\bar{X} - z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \le p \le \bar{X} + z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}\right).
\end{aligned}
$$

So an *approximate* two-sided CI for p is

$$
\boxed{p \in \bar{X} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.}
$$

Similarly, an approximate lower CI is

$$
\bar{X} - z_{\alpha}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \le p,
$$

and an approximate upper CI is

$$
p \le \bar{X} + z_{\alpha}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.
$$

Example: The probability that a student correctly answers a certain test question is p.

Suppose that a random sample of 200 students yields 160 correct answers to the question.

A 95% two-sided CI for p is given by

$$\begin{aligned} p &\in \bar{X} \pm z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \\ &= 0.8 \pm 1.96 \sqrt{\frac{(0.8)(0.2)}{200}} \\ &= 0.8 \pm 0.055. \quad \Box \end{aligned}$$

The half-width of the two-sided CI is

$$z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}}.$$

How many observations should we take so that the half-length is $\leq \epsilon$?

$$z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n}} \;\leq\; \epsilon \;\iff\; n \;\geq\; (z_{\alpha/2}/\epsilon)^2\bar{X}(1-\bar{X}).$$

Of course, $\bar{X}$ is unknown before taking observations. A *conservative* choice for n arises by maximizing $\bar{X}(1-\bar{X}) = 1/4$. Then we have

$$n \;\geq\; z^2_{\alpha/2}/(4\epsilon^2).$$

On the other hand, if we can somehow make a *preliminary estimate* $\hat{p}$ of p (based on a preliminary sample mean), we could use

$$n \;\geq\; z^2_{\alpha/2}\,\hat{p}(1-\hat{p})/\epsilon^2.$$

Example: Suppose we want to take enough observations so that a 95% two-sided CI for p will be no bigger than $\pm 3\%$. How many are required?

$$n \;\geq\; z^2_{\alpha/2}/(4\epsilon^2) \;=\; \frac{(1.96)^2}{4\,(0.03)^2} \;=\; 1067.1.$$

So take $n = 1068$ samples (though this will likely be conservative).

Instead suppose that a pilot study gives an estimate of $\hat{p} = 0.7$. How should n be revised?

$$n \;\geq\; z^2_{\alpha/2}\,\hat{p}(1-\hat{p})/\epsilon^2 \;=\; \frac{(1.96)^2}{(0.03)^2}(0.7)(0.3) \;=\; 896.4.$$

So now we only have to take $n = 897$. $\square$

Georgia Tech

Remark: We can also derive approximate CIs for the **difference between two competing Bernoulli parameters**.

Suppose that $X_1, X_2, \ldots, X_n \stackrel{\text{iid}}{\sim} \text{Bern}(p_x)$ and $Y_1, Y_2, \ldots, Y_m \stackrel{\text{iid}}{\sim} \text{Bern}(p_y)$, where the two samples are independent of each other.

Using the techniques already developed in this lesson, we can easily obtain the following approximate two-sided $100(1-\alpha)\%$ CI for the difference of the success proportions:

$$\boxed{p_x - p_y \ \in \ \bar{X} - \bar{Y} \pm z_{\alpha/2} \sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\bar{Y}(1-\bar{Y})}{m}}.}$$

Example: The probabilities that Georgia Tech and University of Georgia students earn at least a 700 on the Math portion of the SAT test are $p_{☺}$ and $p_{☹}$, respectively.

A random sample of 200 GT students yielded 160 students who scored ≥ 700 on the test. But, sadly, a sample of 400 UGA students revealed only 50 who succeeded.

Then a 95% CI for the difference in success probabilities is

$$\begin{aligned} p_{☺} - p_{☹} &\in \bar{X} - \bar{Y} \pm z_{\alpha/2}\sqrt{\frac{\bar{X}(1-\bar{X})}{n} + \frac{\bar{Y}(1-\bar{Y})}{m}} \\ &= 0.8 - 0.125 \pm 1.96\sqrt{\frac{(0.8)(0.2)}{200} + \frac{(0.125)(0.875)}{400}} \\ &= 0.675 \pm 0.064. \end{aligned}$$

Not looking real good for UGA. $\Box$

Coming Up: Hypothesis Testing!