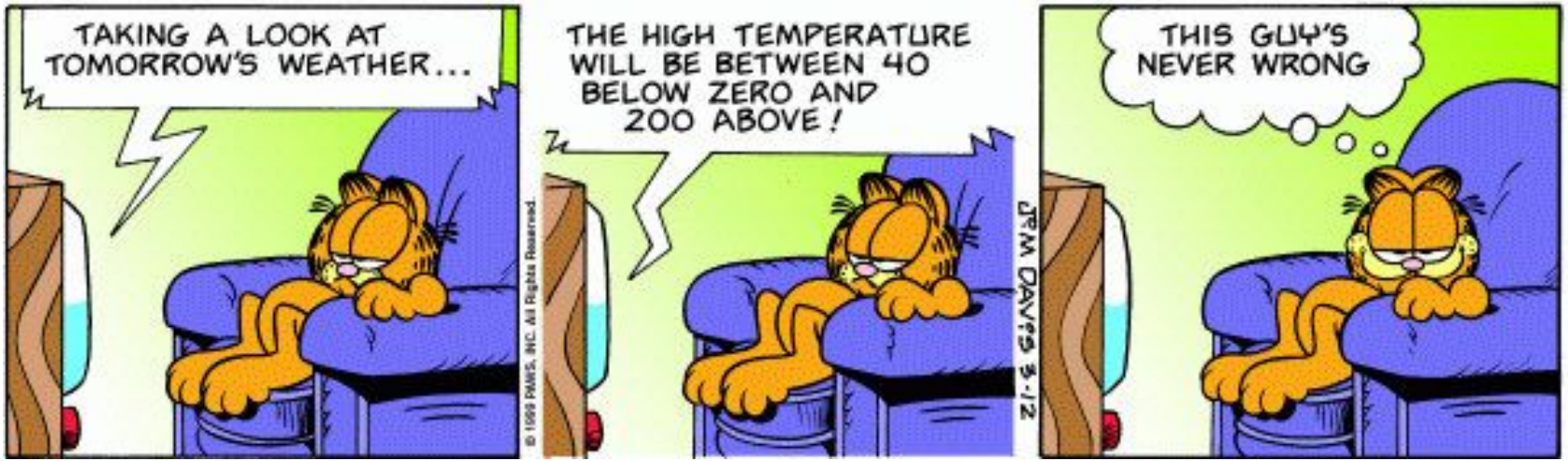
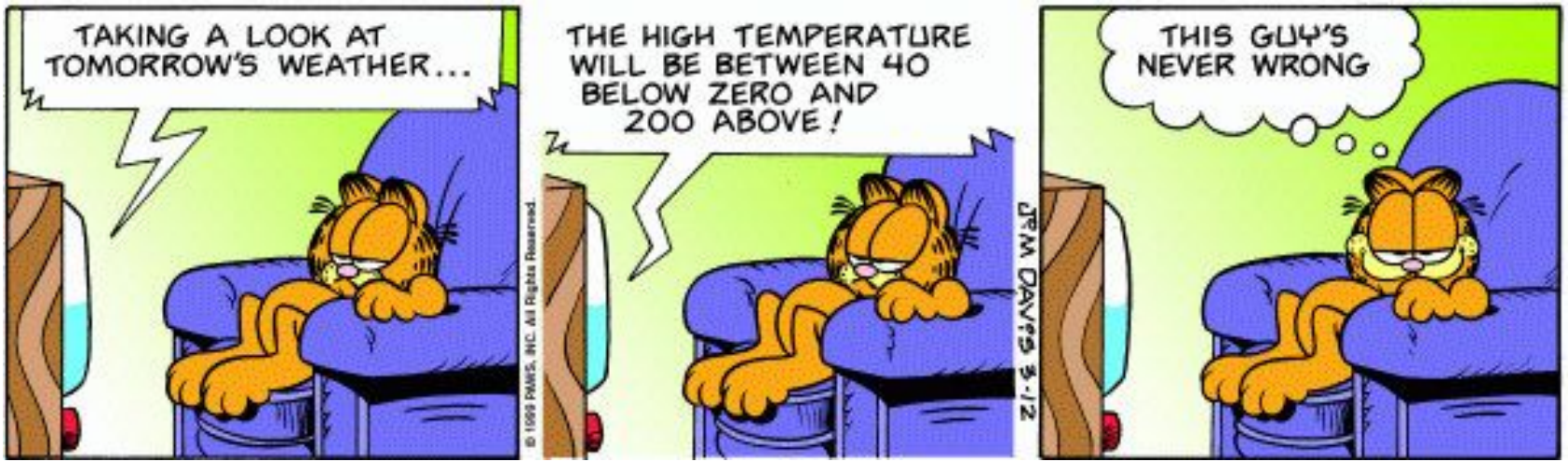
Lecture 14: Intro to Statistical Inference and Confidence Intervals



Announcements and reminders

- HW 3 posted! And due Monday 18 March (2 weeks)
- Quizlet 6 posted! And due Friday at 10 AM



Previously, on CSCI 3022…

Proposition: If X is a normally distributed random variable with mean μ and standard deviation σ, then Z follows a standard normal distribution if we define:

$$Z = \frac{X - \mu}{\sigma} \qquad \text{and} \qquad X = \sigma Z + \mu$$

Fact: If Z is a standard normal random variable, then we can compute probabilities using the standard normal cdf

$$\Phi(z) = P(Z \leq z) = \int_{-\infty}^{z} f(x)\, dx$$

The Central Limit Theorem: Let $X_1, X_2, \ldots, X_n$ be iid draws from some distribution. Then, as n becomes large…

$$\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

Statistical Inference

Goal: Want to extract properties of an underlying population by analyzing sampled data

Questions:

- Is the sample mean $\bar{x}$ a good approximation of the population mean μ?
- Is sample proportion $\hat{p}$ a good approximation of population proportion *p*?
- Is there a statistically significant difference between between the mean of two samples?
- If the answer is **Yes**, how sure are we?
- How much data do we need in order to be **confident** in our conclusion?

Confidence Intervals

The CLT tells us that as the sample size *n* increases, the sample mean of X is close to **normally** distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$

Standardizing the sample mean by first subtracting the expected value and dividing by the standard deviation yields a standard normal random variable

Question: How large does the sample need to be if

- … the population is normally distributed?
- … the population is *not* normally distributed?

Confidence Intervals

The CLT tells us that as the sample size *n* increases, the sample mean of X is close to **normally** distributed with expected value μ and standard deviation $\sigma/\sqrt{n}$

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

Standardizing the sample mean by first subtracting the expected value and dividing by the standard deviation yields a standard normal random variable

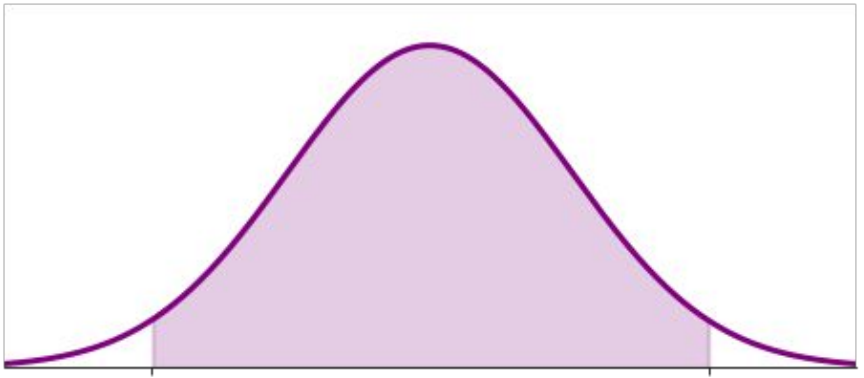
$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

Question: How large does the sample need to be if

- … the population is normally distributed? n ≥ 1
- … the population is *not* normally distributed? n ≥ 30 (roughly)

Confidence Intervals

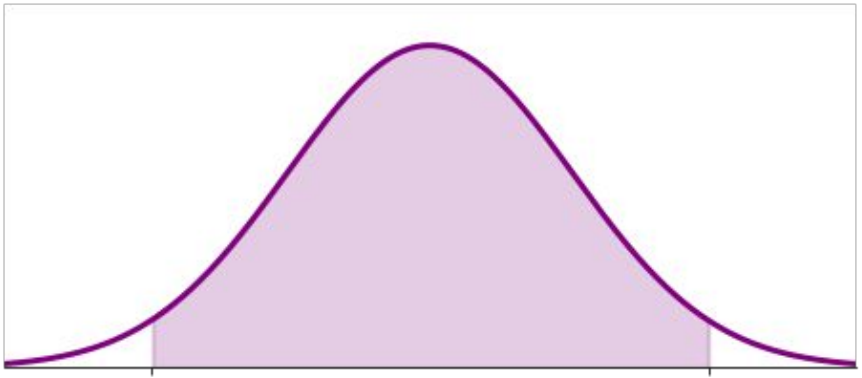
We saw a while ago that 95% of the area under the standard normal curve falls between -1.96 and +1.96, so we know that



Confidence Intervals

We saw a while ago that 95% of the area under the standard normal curve falls between -1.96 and +1.96, so we know that

$P\left(-1.96 \leq Z \leq 1.96\right) = 0.95$ where Z ~ N(0, 1)



… which is equivalent to…

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

So: The **95% confidence interval** is given by the values of X that satisfy that equation!

… how do we find these?

Confidence Intervals

The **95% confidence interval** for the **mean** is given by…

Question: Which things in this expression are **random** and which things are **fixed**?

Confidence Intervals

The **95% confidence interval** for the **mean** is given by…

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

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

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

Question: Which things in this expression are **random** and which things are **fixed**?

Confidence Intervals

The CI varies from sample to sample, so the CI is really a **random interval** itself

Question: S'pose you perform 20 random samples of the population and compute 95% CIs for each sample? How many of the intervals do you expect to contain the true population mean μ?

Confidence Intervals

The CI is centered at _____ and extends _____ to each side of _____

The CI's width is _____ which is **not** random; only the location of the interval's **midpoint** _____ is random

We often write the CI _____ as _____

Interpreting the Confidence Level

Statement: We are 95% confident that the true population mean is in this interval.

What does this even mean?!

Interpreting the Confidence Level

Statement: We are 95% confident that the true population mean is in this interval.

What does this even mean?!

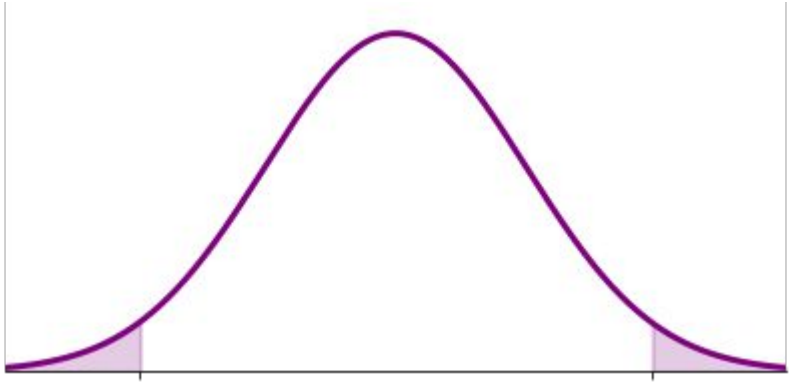
Correct interpretation: In repeated sampling, 95% of all CIs obtained from sampling will actually contain the true population mean. The other 5% of CIs will not.

The confidence level is **not** a statement about any one particular interval.

Instead, it describes what would happen if a very large number of CIs were computed using the same CI formula.

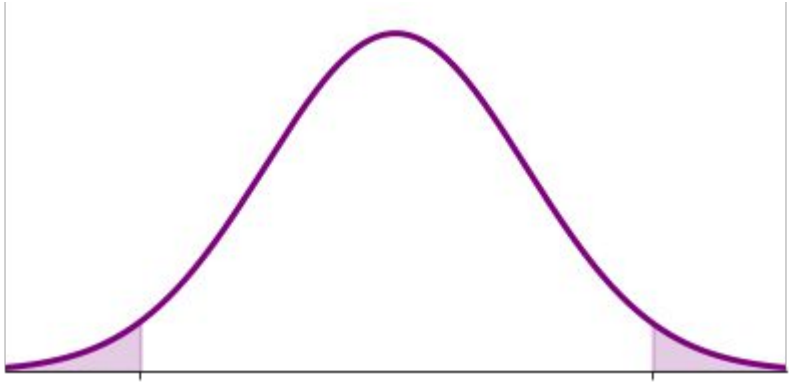
Other levels of confidence

A probability of 1 - α is achieved by using $z_{\alpha/2}$ in place of $z_{0.05/2} = z_{0.025} = 1.96$



Other levels of confidence

A probability of 1 - α is achieved by using $z_{\alpha/2}$ in place of $z_{0.05/2} = z_{0.025} = 1.96$



A 100·(1-α)% confidence interval for the mean μ when the value of σ is known is given by

$$\left[\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}},\quad \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] \quad \text{or} \quad \bar{x} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours/day. Suppose further that the known standard deviation of the characteristic is 2 hours/day.

Find a 90% confidence interval for the amount of relaxation hours/day.

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours/day. Suppose further that the known standard deviation of the characteristic is 2 hours/day.

Find a 90% confidence interval for the amount of relaxation hours/day.

90% → α = 0.1 → $z_{\alpha/2} = z_{0.05}$ = scipy.stats.norm.ppf(0.95) = 1.645

$$\frac{\sigma}{\sqrt{n}} = \frac{2}{\sqrt{1000}} = 0.0632$$

$$\rightarrow \quad \text{CI} = \bar{X} \pm z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = 3.6 \pm 1.645 \cdot \frac{2}{\sqrt{1000}} = 3.6 \pm 1.645 \cdot 0.0632$$

$$= 3.6 \pm 0.104$$

$$\text{CI} = [3.496, 3.704]$$

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours/day. Suppose further that the known standard deviation of the characteristic is 2 hours/day.

Find a 95% confidence interval for the amount of relaxation hours/day.

Confidence Intervals

Example: The General Social Survey (GSS) is a sociological survey used to collect data on demographic characteristics and attitudes of residents of the US. In 2010, the survey collected responses from 1,000 US residents. They found that the average number of hours the respondents had to relax or pursue non-work activities was 3.6 hours/day. Suppose further that the known standard deviation of the characteristic is 2 hours/day.

Find a 95% confidence interval for the amount of relaxation hours/day.

95% $\rightarrow$ $\alpha = 0.05 \rightarrow z_{\alpha/2} = z_{0.025}$ = scipy.stats.norm.ppf(0.975) = 1.96

$$\text{CI} = 3.6 \pm 1.96 \cdot \frac{2}{\sqrt{1000}}$$

$$= 3.6 \pm 0.124 = [3.476, 3.724]$$

Question: What are the advantages/disadvantages of a wider confidence interval?

Interpreting the Confidence Level

Concept check: In the previous example, we found a 95% CI for relaxation time to be [3.48, 3.72]. Which of the following statements are true?

A. 95% of Americans spend 3.48 to 3.72 hours per day relaxing after work.

B. 95% of random samples of 1000 residents will yield CIs that contain the true average number of hours that Americans spend relaxing after work each day.

C. 95% of the time the true average number of hours an American spends relaxing after work is between 3.48 and 3.72 hours/day.

D. We are 95% sure that Americans in this sample spend 3.48 to 3.72 hours/day relaxing after work.

Computing required sample size

Example: For the GSS data, how large would n have to be to get a 95% CI with width at most 0.1?

Computing required sample size

Example: For the GSS data, how large would n have to be to get a 95% CI with width at most 0.1?

$$\text{width} = 2 \cdot 1.96 \frac{2}{\sqrt{n}} \overset{\heartsuit}{=} 0.1$$

Solve for n:

$$\sqrt{n} = \frac{2 \cdot 1.96 \cdot 2}{0.1} = 78.4$$
$$n = (78.4)^2 = 6,146.56$$

→ need n at least 6,147

Unknown variance

In the previous example, we assumed that we knew the population standard deviation σ

Question: How often does this happen in real life?

Unknown variance

In the previous example, we assumed that we knew the population standard deviation σ

Question: How often does this happen in real life? **Pretty much never.**

So what is a data scientist to do??

→ If n is sufficiently large, we use the sample variance (standard deviation) instead

$$\text{CI} = \bar{X} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

→ If n is small, we have to do something else (which we'll talk about later)

Confidence intervals for proportions

Let p denote the proportion of “successes” in a population
(e.g., individuals who graduated from college, compute nodes that do not fail on a given day, etc)

A random sample of n individuals is selected, and X is the number of successes in the sample

Then X can be modeled as a _____ random variable with:

Confidence intervals for proportions

Let p denote the proportion of “successes” in a population
(e.g., individuals who graduated from college, compute nodes that do not fail on a given day, etc)

A random sample of n individuals is selected, and X is the number of successes in the sample

Then X can be modeled as a Binomial random variable with:

$$E[X] = np \quad \text{and} \quad Var(X) = np(1-p)$$

Confidence intervals for proportions

The estimator for p is given by _____

The estimator is approximately normally distributed with

$E[\hat{p}] =$ _____ and $Var(\hat{p}) =$ _____

Standardizing the estimate yields:

Z = _____ ~ _____

This gives us a 100·(1-α)% confidence interval of:

Confidence intervals for proportions

The estimator for p is given by $\hat{p} = \frac{X}{n}$

The estimator is approximately normally distributed with

$$E[\hat{p}] = p \qquad Var(\hat{p}) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

Standardizing the estimate yields:

$$Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

This gives us a 100·(1-α)% confidence interval of:

$$\hat{p} \pm z_{\alpha/2} \sqrt{\frac{p(1-p)}{n}}$$

Confidence intervals for proportions

Example: The EPA considers indoor radon levels about 4 picocuries/liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 of the sampled homes to have indoor radon levels about 4 pCi/L. Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

Confidence intervals for proportions

Example: The EPA considers indoor radon levels about 4 picocuries/liter (pCi/L) of air to be high enough to warrant amelioration efforts. Tests in a sample of 200 homes found 127 of the sampled homes to have indoor radon levels about 4 pCi/L. Calculate the 99% confidence interval for the proportion of homes with indoor radon levels above 4 pCi/L.

99% → α = 0.01 → $z_{\alpha/2} = z_{0.005}$ → stats.norm.ppf(0.995) = 2.57

$$\hat{p} = \frac{127}{200} = 0.635$$

$$\text{CI} = \hat{p} \pm z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{200}}$$

$$= 0.635 \pm 2.57 \cdot \sqrt{\frac{0.635(1-0.635)}{200}}$$

$$= [0.548, \quad 0.722]$$

What just happened?

- **Confidence intervals** happened!
    - The world is uncertain. Our first stab at accounting for this.

- Relationship to **Central Limit Theorem**