Homework 2 Solutions

Charlotte Wickham

2017-10-12

1. Central Limit Theorem Exploration

Using the code from lab as a guide, for each of the following distributions, explore how well the CLT Normal approximation approximates the sampling distribution. You should provide:

- a figure that illustrates the population distribution,
- a series of figures of the simulated sampling distributions for different sample sizes, along with the approximation based on the CLT, and
- a brief summary of your observations

a) A continuous Uniform distribution on the interval $[0, 1]$.
b) A Gamma with shape parameter 2, and scale parameter 2.
c) A Beta(0.5, 0.5) distribution.
d) Some other distribution of your choice.

5 points total. Credit for including all required pieces.

2. Applying the CLT

1pt per part for 5 points total

a) Consider a continuous Uniform(-1, 1) population, and an i.i.d sample of size, n = 25.
   i) Simulate to estimate $P(0.25 < \overline{Y} < 0.75)$ where $\overline{Y}$ is the sample mean.

```
library(tidyverse)
n_sim <- 5000
n <- 25
samples_uni <- rerun(n_sim,
  runif(n = n, min = -1, max = 1))

samples_uni_means <- map_dbl(samples_uni, ~ mean(.x))

mean(samples_uni_means > 0.25 & samples_uni_means < 0.75)
```

```
## [1] 0.0146
```

   $P(0.25 < \overline{Y} < 0.75)$ is estimated to be 0.015
   ii) Use the Central Limit Theorem to approximate the same probability.
   Let $Y \sim Uniform(-1, 1)$, then $\mu = E(Y) = \frac{1}{2}(-1+1) = 0$, and $\sigma^2 = Var(Y) = \frac{1}{12}(1-(-1))^2 = \frac{1}{3}$.
   CLT says,

$$\overline{Y} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(0, \frac{1}{75}\right)$$

   Therefore,

$$\begin{aligned} P(0.25 < \overline{Y} < 0.75) &\approx \texttt{pnorm(0.75, mean = 0, sd = sqrt(1/75)) -} \\ &\quad \texttt{pnorm(0.25, mean = 0, sd = sqrt(1/75))} \\ &= 0.0151914 \end{aligned}$$

   According to the CLT $P(0.25 < \overline{Y} < 0.75)$ is approximately 0.015.

b) Consider a Bernoulli(0.4) population, and an i.i.d sample of size, n = 5.

i) Simulate to estimate $P(\overline{Y} < 0.3)$ where $\overline{Y}$ is the sample mean.

```
n <- 5
samples_bern <- rerun(n_sim,
  rbinom(n = 5, size = 1, prob = 0.4))

samples_bern_means <- map_dbl(samples_bern, ~ mean(.x))

mean(samples_bern_means < 0.3)
```

```
## [1] 0.338
```

$P(\overline{Y} < 0.3)$ is estimated to be 0.34.

ii) Find the same probability exactly using the fact that the sum of n Bernoulli(p) random variables has a Binomial(n, p) distribution and the `pbinom()` function in R.

$$\begin{aligned} P\left(\overline{Y} < 0.3\right) &= P\left(\sum_{i=1}^{5} Y_i < 0.3(5)\right) \\ &= P(X < 1.5) \quad \text{where } X \sim \text{Binomial(5, 0.4)} \\ &= \texttt{pbinom(1.5, size = 5, prob = 0.4)} \\ &= 0.33696 \end{aligned}$$

The exact probability is 0.33696.

iii) Use the Central Limit Theorem to approximate the same probability. Let $Y \sim$ Bernoulli$(p = 0.4)$, then $\mu = E(Y) = p = 0.4$, and $\sigma^2 = Var(Y) = p(1-p) = 0.24$.
CLT says,

$$\overline{Y} \dot{\sim} N\left(\mu, \frac{\sigma^2}{n}\right) = N\left(0.4, \frac{0.24}{5}\right)$$

Therefore,

$$\begin{aligned} P(\overline{Y} < 0.3) &\approx \texttt{pnorm(.3, mean = 0.4, sd = sqrt(.24/5))} \\ &= 0.3240384 \end{aligned}$$

The CLT approximation for this probability is 0.32.

3. Z-test

$n = 25, \overline{Y} = 2.8, \sigma^2 = 1.96$

a. (0.5 points) Compute the Z-statistic for testing the null hypothesis $H_0 : \mu = 2.6$

$$Z = \frac{\overline{Y} - \mu_0}{\sqrt{\sigma^2/n}} = \frac{2.8 - 2.6}{\sqrt{1.96/25}} = \frac{0.2}{0.28} = 0.714$$

b. (0.5 points) Upper one-sided alternative $\implies$ Reject H_0 for $Z > z_{1-\alpha}$.

$z_{1-\alpha} = \texttt{qnorm(0.9)} = 1.282$.

$$Z = 0.714 < 1.282 \implies \text{Fail to reject the null hypothesis}$$

c. (0.5 points) Now suppose that we perform a Z-test, but incorrectly assume that the population variance is 0.49 instead of the true value 1.96. If the null hypothesis is true, what proportion of the time will we reject the null if we are using a level $\alpha = 0.1$ critical value?

Let σ_0^2 be the true value of σ^2, and σ_a^2 be the assumed value, we want

$$\begin{aligned}
P_{H_0}\left(Z(\mu_0) > z_{1-\alpha}\right) &= P_{H_0}\left(\frac{\overline{Y}-\mu_0}{\sqrt{\sigma_a^2/n}} > z_{1-\alpha}\right) \\
&= P_{H_0}\left(\frac{\overline{Y}-\mu_0}{\sqrt{\sigma_0^2/n}} > \frac{\sqrt{\sigma_a^2/n}}{\sqrt{\sigma_0^2/n}} z_{1-\alpha}\right) \quad \text{Multiplying both sides by } \frac{\sqrt{\sigma_a^2/n}}{\sqrt{\sigma_0^2/n}} \\
&\approx P\left(Z > \sqrt{\frac{\sigma_a^2}{\sigma_0^2}} z_{1-\alpha}\right) \quad \text{where } Z \sim N(0,1) \\
&= P\left(Z > 0.5 z_{1-\alpha}\right) \\
&= \texttt{1 - pnorm(0.5 * qnorm(1 - 0.1))} \\
&= 0.2608342
\end{aligned}$$

We will reject the null hypothesis 26% of the time, significantly more than our desired 10% of the time.

d. (0.5 points) If the true population mean household size in Corvallis is 2.7, what is the power of this test?

From lecture

$$\begin{aligned}
P_{\mu_A}(\text{Reject } H_0) &\approx P\left(Z > z_{1-\alpha} - \sqrt{n}\frac{\mu_A - \mu_0}{\sigma}\right) \quad \text{where } Z \sim N(0,1) \\
P\left(Z > z_{1-\alpha} - \sqrt{n}\frac{\mu_A - \mu_0}{\sigma}\right) &= P\left(Z > z_{1-0.1} - \sqrt{25}\frac{2.7-2.6}{1.4}\right) \\
&= P\left(Z > 1.2815516 - 0.3571429\right) \\
&= P\left(Z > 0.9244087\right) \\
&= \texttt{1 - pnorm(0.9244087)} \\
&= 0.1776368
\end{aligned}$$

The power is 0.18.

e. (0.5 points) If instead of n = 25 people in your sample, you had n = 100 people in your sample, what would the power of the test be?

```
n <- 100
```

$$\begin{aligned}
P_{\mu_A}(\text{Reject } H_0) &\approx P\left(Z > z_{1-\alpha} - \sqrt{n}\frac{\mu_A - \mu_0}{\sigma}\right) \quad \text{where } Z \sim N(0,1) \\
P\left(Z > z_{1-\alpha} - \sqrt{n}\frac{\mu_A - \mu_0}{\sigma}\right) &= P\left(Z > z_{1-0.1} - \sqrt{100}\frac{2.7-2.6}{1.4}\right) \\
&= P\left(Z > 1.2815516 - 0.7142857\right) \\
&= P\left(Z > 0.5672659\right) \\
&= \texttt{1 - pnorm(0.5672659)} \\
&= 0.2852668
\end{aligned}$$

The power is 0.29.

f. (0.5 points) If instead of performing a level $\alpha = 0.1$ Z-test with your original $n = 25$ people, you performed a level $\alpha = 0.5$ Z-test with that same sample size, what would the power of the test be?

$$P_{\mu_A}(\text{Reject } H_0) \approx P\left(Z > z_{1-\alpha} - \sqrt{n}\frac{\mu_A - \mu_0}{\sigma}\right) \quad \text{where } Z \sim N(0,1)$$

$$\begin{aligned} P\left(Z > z_{1-\alpha} - \sqrt{n}\frac{\mu_A - \mu_0}{\sigma}\right) &= P\left(Z > z_{1-0.5} - \sqrt{100}\frac{2.7 - 2.6}{1.4}\right) \\ &= P\left(Z > 0 - 0.7142857\right) \\ &= P\left(Z > -0.7142857\right) \\ &= \texttt{1 - pnorm(-0.7142857)} \\ &= 0.7624747 \end{aligned}$$

The power is 0.76.

g. (2 points) Can you generalize these results? How does power change as you increase sample size? How does power change as you increase the significance level?

The larger the sample the higher the power for any specific alternative (part of the consistency of Z-test).

The higher the significance level, α, the higher the power for any specific alternative. Similar justification, but observe that as $\alpha \to 1$ the critical value, $z_{1-\alpha}$, on inequality on RHS $\to -\infty$, and $P\left(Z > -\infty\right) = 1$.

4. Good estimates

(5 points)

For a given sample of size n.

Repeat many, B, times:

1. Sample n i.i.d from a Normal(μ, σ^2). To do this we'll have to pick a specific mean and variance, say `mu` and `sigma`.
2. Find the sample mean of the sample, call it $\overline{Y}^{(k)}$

With the collection of B sample means, $\overline{Y}^{(1)}, \ldots, \overline{Y}^{(B)}$, or in R maybe `sample_means`:

- Estimate $E\left(\overline{Y}\right)$ with their average:

$$\widehat{E\left(\overline{Y}\right)} = \frac{1}{B}\sum_{k=1}^{B}\overline{Y}^{(k)}$$

`mean(sample_means)`

- Estimate $Var\left(\overline{Y}\right)$ with variance of sample means:

$$\widehat{Var\left(\overline{Y}\right)} = \frac{1}{B-1}\sum_{k=1}^{B}\left(\overline{Y}^{(k)} - \widehat{E\left(\overline{Y}\right)}\right)^2$$

`var(sample_means)`

Bias

The simulation based estimate of the Bias is $\widehat{E\left(\overline{Y}\right)} - \mu$, or `mu - mean(sample_means)`

MSE

The simulation based estimate of the MSE is

$$\left(\widehat{E\left(\overline{Y}\right)} - \mu\right)^2 + \widehat{Var\left(\overline{Y}\right)}$$

```
(mu - mean(sample_means))^2 + var(sample_means)
```

Consistency

We could then repeat for increasing values of n. To evaluate consistency we want to see our MSE estimate approaching zero.

Limtations:

1. We need to pick a specific μ and σ to proceed. We can't make a general statement about all μ and σ without analytical justification.
2. While we can increase n, seeing the MSE decreasing isn't the same as proving its limit is zero.
3. Our simulations will always be subject to some error, we may miss important deviations that are below our level of detection.