Statistical inference

Exercises and their answers

Brian Caffo, Jeff Leek, Roger Peng (assembly of exercises and all answers by Henriette Hamer (https://www.sledgehammer-productions.nl))

1. Introduction

1. The goal of statistical inference is to?
   a. Infer facts about a population from a sample.
   b. Infer facts about the sample from a population.
   c. Calculate sample quantities to understand your data.
   d. To torture Data Science students.

   Answer:
   A. Infer facts about a population from a sample.

2. The goal of randomization of a treatment in a randomized trial is to?
   a. It doesn't really do anything.
   b. To obtain a representative sample of subjects from the population of interest.
   c. Balance unobserved covariates that may contaminate the comparison between the treated and control groups.
   d. To add variation to our conclusions.

   Answer:
   C: Balance unobserved covariates that may contaminate the comparison between the treated and control groups.

3. Probability is ?
   a. A population quantity that we can potentially estimate from data.
   b. A data quantity that does not require the idea of a population.

   Answer:
   A. A population quantity that we can potentially estimate from data.

2. Probability

1. Can you add the probabilities of any two events to get the probability of at least one occurring?

   Answer:
   Yes, but you have to take into account the fact that they can occur both.

2. I define a PMF, p so that for $x = 0$ and $x = 1$ we have $p(0) = -0.1$ and $p(1) = 1.1$. Is this a valid PMF?

   $P(A \cup A) = P(A) + P(B) - P(A \cap B)$

   Answer:
   No, for a valid PMF all $p(x) \geq 0$, and $p(0) < 0$

3. What is the probability that 75% or fewer calls get answered in a randomly sampled day from the population distribution from this chapter?

   Answer:

```
# for elaborate explanation, see syllabus text, as it's literaly taken from it
1.5 * 0.75 / 2
```

```
## [1] 0.5625
```

4. The 97.5th percentile of a distribution is?

   Answer:
   The point on the x-axis where the surface below the density curve left of that point is equal to 0.975.
   $F(x) = P(X \leq x) = 0.975$

3. Conditional probability

1. I pull a card from a deck and do not show you the result. I say that the resulting card is a heart. What is the probability that it is the queen of hearts?

   Answer:
   $\frac{1}{13}$.
   $P(A)$:probability of drawing a queen (which is $\frac{4}{52} = \frac{1}{13}$)
   $P(B)$: probability of drawing a heart (which is $\frac{13}{52} = \frac{1}{4}$)
   $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A) = \frac{1}{13}$
   $P(A \cap B) = P(A)P(B)$ is valid because the probability of drawing a queen is independent of the suit, and vice versa.

2. The odds associated with a probability, p , are defined as?

   Answer:
   $\frac{p}{1-p}$

3. $\frac{P(success)}{P(failure)}$ The probability of getting two sixes when rolling a pair of dice is?

   Answer:
   $\frac{1}{36}$, assuming they're both fair dice.
   $P(6) \times P(6) = \frac{1}{6} \times \frac{1}{6} = \frac{1}{36}$

4. Expected values

1. A standard die takes the values 1, 2, 3, 4, 5, 6 with equal probability. What is the expected value?

   Answer:
   3.5
   $\bar{X} = \sum_{i=1}^{n} x_i p(x_i)$ where $p(x_i) = 1/n$
   $E[X] = 1 \times \frac{1}{6} + 2 \times \frac{1}{6} + 3 \times \frac{1}{6} + 4 \times \frac{1}{6} + 5 \times \frac{1}{6} + 6 \times \frac{1}{6} = 3.5$

2. Consider a density that is uniform from -1 to 1. (I.e. has height equal to 1/2 and looks like a box starting at -1 and ending at 1). What is the mean of this distribution?

   Answer:
   0

   ```
   qunif(0.50, min=-1, max=1)
   ```

   ```
   ## [1] 0
   ```

3. If a population has mean μ , what is the mean of the distribution of averages of 20 observations from this distribution?

   Answer:
   μ , for the mean the size of the sample is of no influence.

5. Variation

1. If I have a random sample from a population, the sample variance is an estimate of?
   a. The population standard deviation.
   b. The population variance.
   c. The sample variance.
   d. The sample standard deviation.

Answer:
B: The population variance.

2. The distribution of the sample variance of a random sample from a population is centered at what?
    a. The population variance.
    b. The population mean.

Answer:
A: The population variance.

3. I keep drawing samples of size n from a population with variance σ^2 and taking their average. I do this thousands of times. If I were to take the variance of the collection of averages, about what would it be?

Answer:
$2 \times \frac{\sigma^2}{n}$
I couldn't find the theory behind this! Intuitively I agree, but I don't trust my intuition (in statistics)

4. You get a random sample of n observations from a population and take their average. You would like to estimate the variability of averages of n observations from this population to better understand how precise of an estimate it is. Do you need to repeated collect averages to do this?
    a. No, we can multiply our estimate of the population variance by $1/n$ to get a good estimate of the variability of the average.
    b. Yes, you have to get repeat averages.

Answer:
A. No, we can multiply our estimate of the population variance by $1/n$ to get a good estimate of the variability of the average.
But I'm not really sure.

6. Some common distributions

1. Your friend claims that changing the font to comic sans will result in more ad revenue on your web sites. When presented in random order, 9 pages out of 10 had more revenue when the font was set to comic sans. If it was really a coin flip for these 10 sites, what's the probability of getting 9 or 10 out of 10 with more revenue for the new font?

Answer:
$\binom{10}{9} \times 0.5^9 \times 0.5^{10-9} + \binom{10}{10} \times 0.5^{10} \times 0.5^{10-10}$
assuming a fair coin.

```
choose(10, 9) * 0.5^(9) * 0.5^(10-9) + choose(10, 10) * 0.5^(10) * 0.5^(10-10)
```

```
## [1] 0.01074219
```

```
pbinom(8, size=10, prob=0.5, lower.tail = FALSE)
```

```
## [1] 0.01074219
```

```
# note that you put in 8 and not 9 into the R-function.
```

2. A software company is doing an analysis of documentation errors of their products. They sampled their very large codebase in chunks and found that the number of errors per chunk was approximately normally distributed with a mean of 11 errors and a standard deviation of 2. When randomly selecting a chunk from their codebase, what's the probability of fewer than 5 documentation errors?

Answer:
$\mu = 11$
$\sigma = 2$
$P(x = 5) = \frac{5-\mu}{\sigma} = \frac{5-11}{2} = -3$ (think 3σ)

```
pnorm(5, mean=11, sd=2, lower.tail = TRUE)
```

```
## [1] 0.001349898
```

Here I get lost. Based on "99% of the normal density lies within 3sd" (so 1% outside, on both sides, so on one side 0.5%) I would have expected (1-0.99)/2 = 0.005 as an (approximate) answer. 0.0013 is not really 0.005 …

3. The number of search entries entered at a web site is Poisson at a rate of 9 searches per minute. The site is monitored for 5 minutes. What is the probability of 40 or fewer searches in that time frame?

Answer:
$\lambda = 9$ searches/minute
$t = 5$ minutes

```
lambda <- 9
t <- 5
ppois(40, lambda * t, lower.tail = TRUE)
```

```
## [1] 0.2555451
```

7. Asymptopia

1. I simulate 1,000,000 standard normals. The LLN says that their sample average must be close to?

   Answer:
   0, the mean of the population (which is standard normal, so $\mu = 0$)

2. About what is the probability of getting 45 or fewer heads out 100 flips of a fair coin? (Use the CLT, not the exact binomial calculation).

   Answer:
   With $\frac{\hat{p}-p}{\sqrt{p(1-p)/n}}$,
   $\frac{0.45-0.5}{\sqrt{0.5(1-0.5)/100}} = \frac{-0.05}{\sqrt{0.5^2/100}} = \frac{-0.05}{0.5/\sqrt{100}} = \frac{-0.05}{0.5/10} = \frac{-0.05}{0.05} = -1$,
   think -1σ and then (50-68/2 = 16) 16% lies left of $-\sigma$
   Is this the correct train of thought? for some reference:

```
pbinom(45, prob=0.5, size = 100, lower.tail = TRUE)
```

```
## [1] 0.1841008
```

   ok, this seems close enough…

3. Consider the father.son data. Using the CLT and assuming that the fathers are a random sample from a population of interest, what is a 95% confidence mean height in inches?

   Answer:
   In the population the 95% confidence interval is defined as $\mu \pm 2\sigma$
   Because we're dealing with a sample here, one replaces σ with $\sigma/\sqrt{n}$
   It's, by the way, an example from the syllabus itself except for the transformation to feet …

```
library(UsingR)
data(father.son)
x <- father.son$sheight
n <- length(x)
mean(x) + c(-1, 1) * qnorm(.975) * sd(x) / sqrt(n)
```

```
## [1] 68.51605 68.85209
```

4. The goal of a a confidence interval having coverage 95% is to imply that:
   a. If one were to repeated collect samples and reconstruct the intervals, around 95% percent of them would contain the true mean being estimated.
   b. The probability that the sample mean is in the interval is 95%.

   Answer:

A: If one were to repeated collect samples and reconstruct the intervals, around 95% percent of them would contain the true mean being estimated.

5. The rate of search entries into a web site was 10 per minute when monitoring for an hour. Use R to calculate the exact Poisson interval for the rate of events per minute?

Answer:
$\lambda = 10$ seach entries per minute (this is actually $\hat{\lambda}$)
$t = 1$, just talking about minutes here

The Poisson interval is: $\hat{\lambda} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}$, $Z = 1.645$ for 95% confidence interval.
$10 \pm 1.645\sqrt{10} \approx 10 \pm 5.20$

```
10 + c(-1,1)*qnorm(.95)*sqrt(10)
```

```
## [1]  4.798516 15.201484
```

I feel like I'm not using the information that there was monitoring for an hour. Also, maybe $t = 60$ would be better?
$\lambda = 10$ seach entries per minute (this is actually $\hat{\lambda}$)
$t = 60$

The Poisson interval is: $\hat{\lambda} \pm Z_{1-\alpha/2}\sqrt{\frac{\hat{\lambda}}{t}}$, $Z = 1.645$ for 95% confidence interval.
$10 \pm 1.645\sqrt{10/60} \approx 10 \pm 0.67$

```
10 + c(-1,1)*qnorm(.95)*sqrt(10/60)
```

```
## [1]  9.328491 10.671509
```

8. *t* Confidence intervals

1. For iid Gaussian data, the statistic $\frac{\bar{X}-\mu}{s/\sqrt{n}}$ must follow a:
   a. Z distribution
   b. *t* distribution

   Answer:
   A: Z distribution assuming Z means standard normal. If not, then the answer is perhaps t after all?

2. Paired differences T confidence intervals are useful when:
   a. Pairs of observations are linked, such as when there is subject level matching or in a study with baseline and follow up measurements on all participants.
   b. When there was randomization of a treatment between two independent groups.

Answer:
A: Pairs of observations are linked, such as when there is subject level matching or in a study with baseline and follow up measurements on all participants.

3. The assumption that the variances are equal for the independent group T interval means that:
    a. The sample variances have to be nearly identical.
    b. The population variances are identical, but the sample variances may be different.

Answer:
B: The population variances are identical, but the sample variances may be different.
But I’m not sure, can’t explain why it is so.

9. Hypothesis testing

1. Which hypothesis is typically assumed to be true in hypothesis testing?
    a. The null.
    b. The alternative.

Answer:
A: The null.

2. The type I error rate controls what?

Answer:
“We have fixed α to be low, so if we reject H_0 (either our model is wrong) or there is a low probability that we have made an error.”
The probability that we reject H_0 while it is actually true.

10. P-values

1. P-values are probabilities that are calculated assuming which hypothesis is true?
    a. the alternative
    b. the null

Answer:
B: the null

2. You get a P-value of 0.06. Would you reject for a type I error rate of 0.05?
    a. Yes you would reject the null
    b. No you would not reject the null
    c. It depends on information not given

Answer:
B: No you would not reject the null
H_0 is rejected when $P_{value} < \alpha$, as $0.06 > 0.05$ we fail to reject.

3. The proposed procedure for getting a two sided P-value for the exact binomial test considered here is what?
   a. Multiplying the one sided P-value by one half
   b. Doubling the larger of the two one sided P-values
   c. Doubling the smaller of the two one sided P-values
   d. No procedure exists

Answer:
C: Doubling the smaller of the two one sided P-values

11. Power

1. Power is a probability calculation assuming which is true:
   a. The null hypothesis
   b. The alternative hypothesis
   c. Both the null and alternative

Answer:
A: The null hypothesis
In hypothesis testing it's always about H_0 being true, unless we can prove that it's not. The bigger the power, the more 'prove' that H_0 is not true.

2. As your sample size gets bigger, all else equal, what do you think would happen to power?
   a. It would get larger
   b. It would get smaller
   c. It would stay the same
   d. It cannot be determined from the information given

Answer:
A: It would get larger
I could make a visual when I figure out how to give an area a color in a plot. Hints on how to do that are more than welcome :-)

3. What happens to power as μ_a gets further from μ_0 ?
   a. Power decreases
   b. Power increases

c. Power stays the same
d. Power oscillates

Answer:
B: Power increases

4. In the context of calculating power, the effect size is?
    a. The null mean divided by the standard deviation
    b. The alternative mean divided by the standard error
    c. The difference between the null and alternative means divided by the standard deviation
    d. The standard error divided by the null mean

Answer:
C: The difference between the null and alternative means divided by the standard deviation

12. The bootstrap and resampling

1. The bootstrap uses what to estimate the sampling distribution of a statistic?
    a. The true population distribution
    b. The empirical distribution that puts probability 1/n for each observed data point

Answer:
B: The empirical distribution that puts probability 1/n for each observed data point

2. When performing the bootstrap via Monte Carlo resampling for a data set of size n which is true? Assume that you're going to do 10,000 bootstrap resamples?
    a. You sample n complete data sets of size 10,000 with replacement
    b. You sample 10,000 complete data sets of size n without replacement
    c. You sample 10,000 complete data sets of size n with replacement
    d. You sample n complete data sets of size 10,000 without replacement

Answer:
C: You sample 10,000 complete data sets of size n with replacement
B and D you can reject immediately, as the whole idea is that you use your sample as a substitute population.

3. Permutation tests do what?
    a. Creates a null distribution for a hypothesis test by permuting a predictor variable.
    b. Creates a null distribution by resampling from the response with replacement.
    c. Creates an alternative distribution by permuting group labels.
    d. Creates confidence intervals by resampling with replacement.

Answer:
C: Creates an alternative distribution by permuting group labels.

This document was created with Win2PDF available at http://www.win2pdf.com.
The unregistered version of Win2PDF is for evaluation or non-commercial use only.
This page will not be added after purchasing Win2PDF.