Review for final exam: in-class problems
MIT 18.05 Spring 2022

Probability

- **Counting**
  - Sets
  - Inclusion-exclusion principle
  - Rule of product (multiplication rule)
  - Permutation and combinations
- **Basics**
  - Outcome, sample space, event
  - Discrete, continuous
  - Probability function
  - Conditional probability
  - Independent events
  - Law of total probability
  - Bayes' theorem
- **Random variables**
  - Discrete: general, uniform, Bernoulli, binomial, geometric
  - Continuous: general, uniform, normal, exponential
  - pmf, pdf, cdf
  - Expectation = mean = average value
  - Variance; standard deviation
- **Joint distributions**
  - Joint pmf and pdf
  - Independent random variables
  - Covariance and correlation
- **Central limit theorem**

Statistics

- **Maximum likelihood**
- **Least squares**
- **Bayesian inference**
  - Discrete sets of hypotheses
  - Continuous ranges of hypotheses
  - Beta distributions
  - Conjugate priors
  - Choosing priors
  - Probability intervals
- **Frequentist inference**
  - NHST: rejection regions, significance
  - NHST: p-values
  - z, t, χ^2
  - NHST: type I and type II error
  - NHST: power
  - Confidence intervals
- **Bootstrap confidence intervals**

    - Empirical bootstrap confidence intervals
    - Parametric bootstrap confidence intervals
- **Linear regression**

Problem 1. Basketball
Suppose that against a certain opponent the number of points the MIT basketball team scores is normally distributed with unknown mean θ and unknown variance, σ^2.

Suppose that over the course of the last 10 games between the two teams MIT scored the following points:

$$59, 62, 59, 74, 70, 61, 62, 66, 62, 75$$

(a) Compute a 95% t–confidence interval for θ. Does 95% confidence mean that the probability θ is in the interval you just found is 95%?

(b) Now suppose that you learn that $\sigma^2 = 25$. Compute a 95% z–confidence interval for θ. How does this compare to the interval in (a)?

(c) Let X be the number of points scored in a game. Suppose that your friend is a confirmed Bayesian with *a priori* belief $\theta \sim N(60, 16)$ and that $X \sim N(\theta, 25)$. He computes a 95% probability interval for θ, given the data in part (a). How does this interval compare to the intervals in (a) and (b)?

(d) Which of the three intervals constructed above do you prefer? Why?

Problem 2. Confidence interval 2
The volume in a set of wine bottles is known to follow a N$(\mu, 25)$ distribution. You take a sample of the bottles and measure their volumes. How many bottles do you have to sample to have a 95% confidence interval for μ with width 1?

Problem 3. Polling confidence intervals
You do a poll to see what fraction p of the population supports candidate A over candidate B.

(a) How many people do you need to poll to know p to within 1% with 95% confidence.

(b) Let p be the fraction of the population who prefer candidate A. If you poll 400 people, how many have to prefer candidate A so that the 90% confidence interval is entirely above $p = 0.5$.

Problem 4. Confidence intervals 3
Suppose you made 40 confidence intervals with confidence level 95%. About how many of them would you expect to be "wrong'? That is, how many would not actually contain the parameter being estimated? Should you be surprised if 10 of them are wrong?

Problem 5. (Confidence intervals)
A statistician chooses 20 randomly selected class days and counts the number of students present in 18.05. They find a standard deviation of 4.06 students If the number of students present is normally distributed, find the 95% confidence interval for the population standard deviation of the number of students in attendance.

Problem 6. Linear regression (least squares)

(a) Set up fitting the least squares line through the points $(1, 1)$, $(2, 1)$, and $(3, 3)$.

Problem 7. Empirical bootstrap
Suppose we had 100 data points $x_1, \ldots x_{100}$ with sample median $\widehat{q_{0.5}} = 3.3$.

(a) Outline the steps needed to generate an empirical percentile bootstrap 90% confidence interval for the median $q_{0.5}$.

(b) Suppose now that the sorted list in the previous problem consists of 200 empirical bootstrap medians computed from resamples of size 100 drawn from the original data. Use the list to construct a 90% precentile CI for $q_{0.5}$.

Problem 8. Parametric bootstrap
Suppose we have a sample of size 100 drawn from a geom(p) distribution with unknown p. The MLE estimate for p is given by by $\hat{p} = 1/\bar{x}$. Assume for our data $\bar{x} = 3.30$, so $\hat{p} = 1/\bar{x} = 0.30303$.

(a) Outline the steps needed to generate a parametric basic bootstrap 90% confidence interval.

(b) Suppose the following sorted list consists of 200 bootstrap means computed from a sample of size 100 drawn from a geometric(0.30303) distribution. Use the list to construct a 90% basic CI for p.

```
2.68 2.77 2.79 2.81 2.82 2.84 2.84 2.85 2.88 2.89
2.91 2.91 2.91 2.92 2.94 2.94 2.95 2.97 2.97 2.99
3.00 3.00 3.01 3.01 3.01 3.03 3.04 3.04 3.04 3.04
3.04 3.05 3.06 3.06 3.07 3.07 3.07 3.08 3.08 3.08
3.08 3.09 3.09 3.10 3.11 3.11 3.12 3.13 3.13 3.13
3.13 3.15 3.15 3.15 3.16 3.16 3.16 3.16 3.17 3.17
3.17 3.18 3.20 3.20 3.20 3.21 3.21 3.22 3.23 3.23
3.23 3.23 3.23 3.24 3.24 3.24 3.24 3.25 3.25 3.25
3.25 3.25 3.25 3.26 3.26 3.26 3.26 3.27 3.27 3.27
3.28 3.29 3.29 3.30 3.30 3.30 3.30 3.30 3.30 3.31
3.31 3.32 3.32 3.34 3.34 3.34 3.34 3.35 3.35 3.35
3.35 3.35 3.36 3.36 3.37 3.37 3.37 3.37 3.37 3.37
3.38 3.38 3.39 3.39 3.40 3.40 3.40 3.40 3.41 3.42
3.42 3.42 3.43 3.43 3.43 3.43 3.44 3.44 3.44 3.44
3.44 3.45 3.45 3.45 3.45 3.45 3.45 3.45 3.46 3.46
3.46 3.46 3.47 3.47 3.49 3.49 3.49 3.49 3.49 3.50
3.50 3.50 3.52 3.52 3.52 3.52 3.53 3.54 3.54 3.54
3.55 3.56 3.57 3.58 3.59 3.59 3.60 3.61 3.61 3.61
3.62 3.63 3.65 3.65 3.67 3.67 3.68 3.70 3.72 3.72
3.73 3.73 3.74 3.76 3.78 3.79 3.80 3.86 3.89 3.91
```

Problem 9. (NHST chi-square)
A study of recidivism (repeat offenses) of juvenile offenders used an experimental design with random assignment of juveniles to experimental intervention (Family Group Counseling) or

control group (diversion programs). 70 out of 200 people in the control group re-offended and 30 out of 200 people in the experimental group re-offended.

Use a chi-square significance test to test whether the recidivism rates within 6 months for the two experimental groups are significantly different at a significance level of 0.05.