# FINAL EXAM REVIEW (NEW TOPICS)

Stat250 S25

Prof Amanda Luby

---

The final exam will take place on **Saturday, June 7 from 3:30-6pm**. The exam is closed book, but you can use the formula sheet provided by me along with both sides of one notecard of your own notes. You will also have access to a calculator.

To study, I recommend carefully going through class notes, homework problems, previous exams and exam prep, and this handout. After reviewing those materials, I recommend solving lots and lots of practice problems (e.g. the problems in the textbook with answers in the back).

This document focuses on the course content since Exam II, but all material from the course is "fair game" for the final exam.

## 1. $\chi^2$ tests

- Know when to use chi-squared goodness of fit vs test of independence
- Know assumptions of the chi-squared test (e.g., expected counts $\geq 5$)
- Write appropriate null and alternative hypotheses
- Calculate test statistics and interpret R output

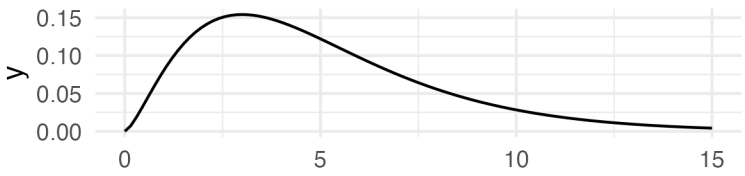
### Goodness of Fit

- Compute expected frequencies
- Compute test statistic using observed and expected frequencies
- Know degrees of freedom for a GOF test
- Provide appropriate R code to find the p-value

**Example question**: A die is rolled 60 times. The observed counts for each face are:

| Face | 1 | 2 | 3 | 4 | 5 | 6 |
|---|---|---|---|---|---|---|
| Count | 8 | 10 | 9 | 11 | 12 | 10 |

(a) State the hypotheses for testing whether the die is fair.
(b) Calculate the expected counts under the null.
(c) Calculate the chi-squared test statistic and degrees of freedom.
(d) Sketch the p-value on the following curve:



### Test of independence

- Compute expected frequencies
- Compute test statistic using observed and expected frequencies
- Know degrees of freedom for a GOF test
- Provide appropriate R code to find the p-value

**Example question**: A random sample of 500 people is surveyed about their smoking and exercise habits. The data and analysis code and results are shown below

| | Smokes | Doesn't Smoke | Total |
|---|---|---|---|
| Exercises | 40 | 180 | 220 |
| Doesn't Ex. | 90 | 190 | 280 |
| Total | 130 | 370 | 500 |

```
chisq.test(smoke_exercise_table)
>
>   Pearson's Chi-squared test
>
> data:  smoke_exercise_table
> X-squared = 12.481, df = 4, p-value = 0.01411
```

(a) What are the null and alternative hypotheses?
(b) Show or explain how the expected counts were computed
(c) Interpret the results of the hypothesis test in context. What doy ou conclude?

## 2. Inference for Regression

- Know the form of a linear regression model and its assumptions
- Interpret coefficients in context
- Use regression output to run hypothesis tests or compute confidence intervals
- Know the difference between a confidence and prediction interval

### Reading R output

- Understand the meaning of coefficients, standard errors, and p-values
- Interpret R's output for `lm()` and `predict()`

**Example Question**: We are interested in prediction penguins' `bill_length` using their `bill_depth`. We fit the following model:

```
Call:
lm(formula = bill_length_mm ~ bill_depth_mm, data = penguins)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-12.8949  -3.9042  -0.3772   3.6800  15.5798

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    55.0674     2.5160  21.887  < 2e-16 ***
bill_depth_mm  -0.6498     0.1457  -4.459 1.12e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.314 on 340 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.05525,   Adjusted R-squared:  0.05247
F-statistic: 19.88 on 1 and 340 DF,  p-value: 1.12e-05
```

(a) Interpret the slope.
(b) Write the regression equation with estimates included.
(c) What is the 95% CI for the slope?

### Theoretical Results

- Manipulate linear regression quantities
- Derive theoretical properties of linear regression estimators

**Example Question**: Let $(x_1, Y_1), (x_2, Y_2), ..., (x_n, Y_n)$ satisfy the conditions for the SLR model.

(a) Write out the SLR model formally. Your model should include $Y_i$, $x_i$, $\beta_0$, $\beta_1$ and $\sigma^2$.
(b) Show that $E(\bar{Y}) = \beta_0 + \beta_1 \bar{X}$
(c) Show that $\text{Var}(\bar{Y}) = \frac{\sigma^2}{n}$

## 3. Bayesian Inference for Beta-Binomial Model

- Know the relationship between Beta and Binomial
- Interpret parameters of the prior, likelihood, and posterior
- Compute posterior summaries (mean, variance, credible intervals)
- Compare Bayesian vs. frequentist perspectives

### Finding the Bayesian estimates

**Example Question**: A baseball player had 18 hits in 60 at-bats We are interested in estimating their "true" batting average (the probability they get a hit in a given at-bat) with uncertainty.

(a) Write out the likelihood function for this data
(b) Suppose your prior belief about the player's true batting average is Beta(3, 7). Find the posterior distribution.

(c) What is the posterior mean?
(d) Provide R code to find the 95% credible interval (the answer is shown below if you'd like to check your answer)

```
[1] 0.1992477 0.4115751
```

### Bayesian Interpretation

- Interpret probability statements about parameters
- Understand how prior beliefs influence the posterior, especially with small sample sizes

**Example**: If someone says "There is a 95% chance that the true batting average is between 0.25 and 0.35," are they using a frequentist or Bayesian interpretation? How can you tell?

**Example**: If we instead used a Unif(0,1) prior in part (b), would the credible interval in (d) be wider, narrower, or stay the same? Explain.