List of Formulas, Procedures and Boxes for the Final Exam, Math 125

(The final will be on Friday, May 20, and covers material from Chapters 3,5,10,12–15,18,21,23, 26,and 28.)

Math 125 *Kovitz* Spring 2022

From Text

Boxes on pages 32, 37, 40 and 41.

Box on page 79.

For example: 2.3 SDs above average says that the standard units are $+2.3$*;*
0.84 SDs below average means that the standard units are -0.84.

Also, when the standard units are 1.36, the value is 1.36 SDs above average;
when the standard units are -2.84*, the value is 2.84 SDs below average.*

Finally, standard units of 0 means that the value is at the average.

Procedure of Example 7 on page 84.

Procedure of Example 9 on page 87.

Boxes on page 160.

Box on page 169.

Box on page 204.

Second box on page 223.

Boxes on pages 229, 230, and 232.

Boxes on pages 241, 242, and 250.

Box on page 259.

Box on page 298.

Box on page 301.

Boxes on pages 325 and 326.

Technical note on page 362.

The first summary point on page 373.

Box on page 378.

Bottom two bullets on page 381.

Box on page 410.

Top box on page 412.

Technical note (ii) on the top of page 415.

Box on page 416.

Procedure for confidence intervals for the population average:
last paragraph on page 416 to the top paragraph on 418.

Boxes on pages 477, 479, 480, and 481.

Section 4 on page 482.

Box on page 527.

Formulas

Drawing a Histogram

First convert the distribution of the population into percentages for each category.

In a histogram the areas of the blocks represent percentages.

> **To figure out the height of a block over a class interval, divide the percent by the length of the interval.**

In a histogram, the height of a block represents crowding—percentage per horizontal unit.

> **With the density scale on the vertical axis, the areas of the blocks come out in percent. The area under the histogram over an interval equals the percentage of cases in that interval. The total area is 100%.**

Conversion to Standard Units

A value is converted to standard units by seeing how many SDs it is above or below the average.

The formula is: $\text{standard units} = \dfrac{\text{observation} - \text{average}}{\text{SD}}.$

Finding Areas under the Normal Curve

An area from minus a value to plus the same value is read off from the Normal Table; other areas are found by making a sketch and expressing the desired area in terms of areas that may be found by using the Table.

The Normal Approximation for Data

If a histogram follows the normal curve, approximate areas may be found by converting the endpoints to standard units and finding the appropriate areas under the normal curve by using the Table.

Regression

The regression line is to a scatter diagram as the average is to a list. The regression line for y on x estimates the average value for y corresponding to each value of x.

Associated with each increase of one SD in x there is an increase of only r SDs in y, on the average. (This way of using the correlation coefficient to estimate the average value of y for each value of x is called the *regression method.*)

A formula for the estimate of the average y for a given x is

$$\text{estimated average} = \text{average } y + (x \text{ in standard units} \times r \times \text{SD of } y)$$

The Regression Fallacy

In virtually all test-retest situations, the bottom group on the first test will on average show some improvement on the second test—and the top group will on average fall back. This is the *regression effect*.

The regression fallacy consists of thinking that the regression effect must be due to something important, not just the spread around the line. The uncertainty of the estimate of the regression line of y on x will lead to more uncertainty when that estimate is subsequently used to estimate x from y using the regression line of x on y; thus the result will not be the original x.

The Regression Line: Slope and Intercept

Associated with each unit increase in x there is some average change in y. The slope of the regression line says how much this change is. The formula for the slope (of the regression line for y on x) is

$$\frac{r \times \text{SD of } y}{\text{SD of } x}$$

To find the intercept of the regression line after the slope m has been found, take any known point (x, y) on the line, substitute it into the equation $y = mx + b$, and then solve for b. The point of averages, which is always on the regression line, is a good point to select for that purpose.

The Chance of the Opposite

The chance of something equals 100% minus the chance of the opposite thing.

The Chance that Both of Two Things Will Happen

Special case when A and B are independent (meaning that P(B|A) = P(B)):

$$\text{P(A and B)} = \text{P(A)P(B)}$$

The Chance that at Least One of Two Things Will Happen

(If you want to find the chance that at least one event occurs, and the events are not mutually exclusive, do not add the chances: the sum will be too big. Blindly adding the chances can give the wrong answer, by double-counting the chance that both things happen. With mutually exclusive events, there is no double-counting.)

Procedure when the two things are not mutually exclusive:

- The opposite thing is that none of those things will happen. This, the opposite thing, means that each of the original things did *not* happen. Its chance can now be calculated by repeated application of the multiplication rule.
- The original chance that we desired—namely the chance that at least one of several things will happen—can now be found by subtracting the chance of its opposite from 100%.

The Binomial Formula

The chance that an event will occur exactly k times out of n is given by the binomial formula

$$\frac{n!}{k!(n-k)!}p^k(1-p)^{n-k}$$

In this formula, n is the number of trials, k is the number of times the event is to occur, and p is the probability that the event will occur on any particular trial. The assumptions:

- The value of n must be fixed in advance.
- p must be the same from trial to trial.
- The trials must be independent.

The first factor in the binomial formula is the binomial coefficient.

The Expected Value for the Sum of the Draws

The expected value for the sum of draws made at random with replacement from a box equals

$$\text{(number of draws)} \times \text{(average of box)}.$$

The Standard Error for the Sum of the Draws

A sum is likely to be around its expected value, but to be off by a chance error similar in size to the standard error.

When drawing at random with replacement from a box of numbered tickets, the standard error for the sum of the draws is

$$\sqrt{\text{number of draws}} \times \text{(SD of box)}.$$

A Shortcut for the SD of a Box with Only Two Kinds of Tickets

When the tickets in the box show only two different numbers, the SD of the box equals

$$\left(\begin{matrix}\text{bigger}\\ \text{number}\end{matrix} - \begin{matrix}\text{smaller}\\ \text{number}\end{matrix}\right) \times \sqrt{\begin{matrix}\text{fraction with}\\ \text{bigger number}\end{matrix} \times \begin{matrix}\text{fraction with}\\ \text{smaller number}\end{matrix}}$$

Classifying and Counting

If you have to classify and count the draws, put 0's and 1's on the tickets. Mark 1 on the tickets that count for you, 0 on the others.

Then find the average of the box and the expected value, and the SD (using the shortcut) and the standard error. Now convert any given values of the count to standard units and use the normal curve to approximate any chance being sought.

The Normal Approximation to Binomial Probabilities

If a binomial probability is considered as the sum of repeated draws from a suitable counting box, the normal approximation may be used—provided the number of trials (draws from the box) is suitably large.

The expected value is the product of the number of trials and the average of the counting box. The standard error is the product of the square root of the number of trials and the SD of the counting box (found by the short cut).

Since the sum of the draws is discrete, it is more accurate to correct the endpoints of the intervals by plus or minus one half.

Next convert the endpoints of the given range to standard units using the formula

$$\textbf{standard units} = \frac{\textbf{given value (corrected)} - \textbf{expected value}}{\textbf{standard error}}.$$

The area under the normal curve between the standard units for the corrected endpoints of the given range will be an approximation for the desired chance.

The Standard Error for a Sample Percentage

$$\textbf{SE for percentage} = \left(\textbf{SD of box}/\sqrt{\textbf{number of draws}}\right)\times\textbf{100\%}.$$

It is a fact that the SD of a 0–1 counting box is always .5 or less. This means that the SE for a percentage is always less than $\dfrac{50\%}{\sqrt{\text{number of draws}}}$, no matter what the percentage of the box.

Warning: if the problem involves classifying and counting to get a percent, put 0's and 1's in the box.

Inference from Sample to Population (the Bootstrap Method)

The bootstrap. When sampling from a 0-1 box whose composition is unknown, the SD of the box can be estimated by substituting the fractions of 0's and 1's in the sample for the unknown fractions in the box. The estimate is good when the sample is reasonably large.

Confidence intervals

A confidence interval for a percentage—with a confidence level specified as a percent—is a range of percentages such that you are that percent confident that the population percentage is in that interval.

A confidence interval is based on the results of a single sample of a particular size.

- the interval "sample percentage ± 2 SEs" is an approximate 95%-confidence interval for the population percentage.
- the interval "sample percentage ± 3 SEs" is an approximate 99.7%-confidence interval for the population percentage.

The EV and SE for the Average of the Draws

When drawing at random from a box:

$$\text{EV for average of draws} = \text{average of box}.$$

$$\text{SE for average of draws} = \frac{\text{SE for sum}}{\text{number of draws}}.$$

If you wish, you may compute the SE for an average directly from the SD of the box by the formula

$$\text{SE for average} = (\text{SD of box})/\sqrt{\text{number of draws}}.$$

Using Normal Curve to Figure Chances for Average of Draws

When drawing at random from the box, the probability histogram for the average of the draws will follow the normal curve, even if the contents of the box do not, The histogram must be put into standard units, and the number of draws must be reasonably large.

Use of the Bootstrap to Estimate the SD of the Box

With a simple random sample, the SD of the sample can be used to estimate the SD of the box. This is helpful when calculating confidence intervals for the population average. The estimate is good when the sample is large.

Various Standard Errors

For a given box model there are several SEs, each showing the likely size of a certain chance error. The corresponding formulas are:

$$\begin{aligned}
\text{SE for sum} &= \sqrt{\text{number of draws}} \times \text{SD of box} \\
\text{SE for average} &= \frac{\text{SE for sum}}{\text{number of draws}} = \frac{\text{SD of box}}{\sqrt{\text{number of draws}}} \\
\text{SE for count} &= \text{SE for sum, from a 0–1 box} \\
\text{SE for percent} &= \frac{\text{SE for count}}{\text{number of draws}} \times 100\% = \frac{\text{SD of zero-one box}}{\sqrt{\text{number of draws}}} \times 100\%
\end{aligned}$$

The SE for the sum is basic; the other formulas all come from that one. These formulas apply to draws made at random with replacement from a box.

Do not confuse the SD and the SE for the average.

- The SD says how far a number in the box is from average—for a typical number.
- The SE for the average says how far the sample average is from the population average—for a typical sample.

Tests of Significance: The Null and the Alternative

To make a test of significance, the null hypothesis has to be formulated as a statement about a box model. Usually, the alternative does too.

- The *null hypothesis* says that an observed difference just reflects chance variation.
- The *alternative hypothesis* says that the observed difference is real.

> The null hypothesis expresses the idea that an observed difference is due to chance. To make a test of signficance, the null hypothesis has to be set up as a box model for the data. The alternative hypothesis is another statement about the box; it says that the difference is real.

Testing Hypotheses: The Observed Significance Level

A test statistic is used to measure the difference between the data and what is expected on the null hypothesis.

z says how many SEs away an observed value is from its expected value, where the expected value is calculated using the null hypothesis. The formula is

$$z = \frac{\text{observed value} - \text{expected value}}{\text{standard error}}.$$

> The observed significance level—often called the P-value—is the chance of getting a test statistic as extreme as or more extreme than the observed one. The chance is computed on the basis that the null hypothesis is right. The smaller the chance is, the stronger the evidence against the null.
>
> > A small value of P indicates that an explanation saying that this is chance variation is unreasonable. We cannot accept the model stated in the null.
> >
> > A large value of P could very well be due to chance variation and we accept the null as a reasonable model.
>
> The P-value of a test is the chance of getting a big test statistic—assuming the null hypothesis to be right. P is not the chance of the null hypothesis being right.

Making a Test of Significance

Based on some available data, the investigator has to—

- translate the null hypothesis into a box model for the data;
- define a test statisic to measure the difference between the data and what is expected on the null hypothesis;
- compute the observed significance level P.

The choice of test statistic depends on the model and the hypothesis being considered.

Making the Decision

Many statisticians have a dividing line that indicates how small the observed significance level has to be before an investigator should reject the null hypothesis.

- If P is less than 5%, the result is called *statistically significant* and the null hypothesis is rejected.

- If P is greater than 5%, we accept the null and state that chance error is a reasoanable explanation for this result.

Tests of Significance Using Zero-one Boxes (Chapter 26, section 5)

If a problem involves classifying and counting, the z-test can be used. The box will contain 0's and 1's.

Once the null hypothesis has been translated into a box model, it is easy to test using

$$z = \frac{\text{observed} - \text{expected}}{\text{SE}}$$

The Chi-Square Test (χ^2-Test)

The χ^2-test is used to check whether a box model for classification involving more than two categories is appropriate in view of certain observed data. This is an approximation to the actual probabilities and may be trusted when each expected frequency in the table is 5 or more.

The χ^2-statistic is obtained by evaluating

$$\chi^2 \;=\; \text{sum of} \;\; \frac{(\text{observed frequency} - \text{expected frequency})^2}{\text{expected frequency}}$$

There is one term in the sum for each line in the table listing observed and expected frequencies.

This statistic measures the distance between observed and expected frequencies.

For the χ^2-test, P is approximately equal to the area to the right of the observed value for the χ^2-statistic, under the χ^2-curve with the appropriate number of degrees of freedom.

When the model is fully specified (no parameters to estimate),

$$\text{degrees of freedom} \;=\; \text{number of terms in } \chi^2 \;-\; \text{one}.$$

- The χ^2-test says whether the data are like the result of drawing at random from a box whose contents are given.

- The z-test says whether the data are like the result of drawing at random from a box whose average is given.

The ingredients of the χ^2-test are the basic data, chance model, frequency table, χ^2-statistic, degrees of freedom, and the observed significance level.