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

(The final will be on Friday, May 17, and covers material from Chapters 3–5,8–18,20,21,23,24,26, and 28.)
Math 125 *Kovitz* Spring 2024

From Text

Boxes on pages 32, 37, 40 and 41.

Endpoint Convention for Rounded Data: If the blocks are designated by whole numbers, each block will start at 1/2 less than the class mark and continue to 1/2 more than the class mark. For example: 67-inch-tall men has a block running from $67 - 1/2$ to $67 + 1/2$, or 66.5 inches to 67.5 inches.

Endpoint Convention for Discrete Counting Data: Here each number has to have an interval one wide centered at that value. For that reason, the base is always $\pm 1/2$ from the number itself. For example: 3 children in family size daa will have an interval running from 2.5 to 3.5 children (see page 43 and the histogram on page 44).

Rule to Find the Area of a Block in a Hisogram: To find the area (or percent of the data) for a block, multiply the length of the interval by the density. Units times % per unit will equal percent.

Types of Variables; Qualitative (descriptive) or quantitative (numeric); continuous (takes on all values in the interval) or discrete (takes on only certain, separated values).

Density Scale: The vertical scale for a histogram. It gives the percent of the area of the histogram per unit of the horizontal scale. It is a density scale, meaning the taller the block of the histogram, the more densely packed it is.

Boxes on pages 59 and 71.

Summary notes 1, 2, 4–8 on pages 76–7. **Note 7 is extremely important for the course.**

Problem 1 on page 77.

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 4 on page 83.

Procedure of Examples 8 and 9 on pages 85 to 87.

Example 10 on pages 90 and 91.

Boxes on pages 113 and 115.

Box on page 132.

Example 1 on pages 132 and 133.

(Optional) Technical note on page 134. This method is allowed on tests and the final exam.

Box on page 150, and text just above the box.

Boxes on page 160.

Box on page 169.

Boxes on page 186.

Top part of box on page 204, but don't use the $x = 0$ method.
Instead find the point of averages and plug it into the equation $y = mx + b$ to get b that way.
(The slope just found is m.)

Second box on page 223.

Boxes on pages 229, 230, 231, 232, and 234.

Boxes on pages 241 (top box only), 242, and 250.

Box on page 259.

Figures 1 and 2 on pages 275 and 276.

Page 276: last *Assistant* comment, continues to page 277 to *Kerrich*.

Boxes on pages 289 and 291, and the top box on page 292..

Box on page 298.

Box on page 301.

Example 1(a) on page 317.

Boxes on pages 325 and 326.

Boxes on pages 359, 360, and 364.

Technical note on page 362.

Box on page 378.

Middle bullet on page 381.

Box on page 386.

Top half of the box on page 410.

Top box on page 412.

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

Problems 10(e) and 10(f) on page 428.

Discussion on pages 475 to 478.

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

Section 26.4 on page 482.

Box on page 547.

Example 1 on page 547.

Text and example 2 from the bottom of page 547 to the middle of page 550.

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%.**

The Average: a measurement of the center.

- The average of a list of numbers equals their sum, divided by how many there are.

Calculation of the standard deviation.

- Find the average.
- Find the deviations from average by subtracting the average from each entry.
- Find the root-mean-square of the deviations from average (by taking their squares, the average of these squares, and the square root of that average—be sure to do these steps in this exact order).

> The Root-mean-square: Square all entries, take the mean (average) of the squares, and take the square root of the mean.
>
> > The term mean-square refers to a square of which we have taken the mean. The term root-mean-square refers to a mean-square of which we have taken the root. That means that the operations are done in reverse order of their positions in the word.
>
> As a formula: r.m.s. size of a list $= \sqrt{\text{average of (entries}^2\text{)}}$.

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.

Percentiles and the Normal Curve

A. For a percentile rank above 50%, subtract 50% from the rank and double the result. That will give you the area between $-z$ and z. Use the normal table to detemine the z's. Here the positive answer is the relevant one.

B. Given a positive z, to find the associated percentile rank, take half of the area between $-z$ and z and add to it the area below $z = 0$ (50%).

C. For a percentile rank below 50%, to find the associated z, double the rank and subtract from 100%. From the table get $-z$ to z for that answer. Choose the negative one here.

D. Given a negative z, to find the associated percentile rank, look up $-z$ to z, subtract that middle area from 100%, and then divide by 2 to get the tail area. The tail area of the left-hand tail gives the percentile rank.

Examples: (a) Percentile rank is 91%. $(91\% - 50\%) \times 2 = 82\%$ z is 1.35.

(b) $z = 0.20$. Half of area is $15.85\%/2 \approx 8\%$. $8\% + 50\% = 58\%$.

(c) Percentile rank is 5%. $100\% - (2 \times 5\%) = 100\% - 10\% = 90\%$. z is -1.65.

(d) $z = -0.50$. -0.50 to 0.50 is about 38% on the table. Subtract $100\% - 38\% = 62\%$. The tail area of the left-hand tail is $62\%/2 = 31\%$. That is the perecntile rank.

Computing the Correlation Coefficient

- Convert the x- and y-values to standard units by finding the average and SD for both the x-values and y-values and then subtracting the average from each value and dividing by the SD. Put the results in two new columns of the table of data.
- For each row multiply (x in standard units) × (y in standard units). Put the resulting products in the last column of the table.
- Take the average of the products. That average is r.

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 predicted y for a given x is

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

The Regression Effect

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 R.M.S. Error for Regression

The r.m.s. error for the regression of y on x can be figured as

$$\sqrt{1-r^2} \times \text{the SD of } y.$$

The units for the r.m.s. error are the same as the units for the variable being predicted.

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 Multiplication Rule

The chance that two things will both happen equals the chance that the first will happen, multiplied by the chance that the second will happen given that the first has happened. $\mathbf{P(A \text{ and } B) = P(A) \times P(B|A)}$.

Independence

Two things are *independent* if the chances for the second given the first are the same, no matter how the first one turns out. Otherwise, the two things are *dependent.*

The Chance that Both of Two Independent Things Will Happen

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

This may be stated verbally:

"If two events are independent, the chance that both will happen equals the product of their unconditional probablilites. This is a special case of the multiplication rule."

There are two important points here.

- Conditional probabilities are not needed when the events are independent.
- The chance being calculated is that both A *and* B will happen. The conjunction is '*and*'.

Mutually Exclusive: Definition

Two things are *mutually exclusive* when the occurrence of one prevents the occurrence of the other: one excludes the other.

The Addition Rule

To find the chance that at least one of two things will happen, check to see if they are mutually exclusive. If they are, add the chances.

The Chance that At Least One of Two Things Will Happen

(The Chance of A or B, when they are Not Mutually Exclusive)

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.

There are several methods to find the chance of either of two events when the events are not mutually exclusive. One is the method of the Chevalier de Méré, presented below as the "Procedure when the things are not mutually exclusive." When applying that procedure, it is good to know that the opposite of 'A or B' is 'not A and not B,' or—as otherwise stated—the opposite of 'either A or B' is 'neither A nor B.'

The Chance that at Least One of Several 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 chances where two or more of the things happen together. With mutually exclusive events, there is no double-counting.)

Procedure when two or more 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. (With 2 things, only 1 application is needed.)
- 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%.

Summary of de Méré procedure: A) opposite of "At least one" (same as A or B or C) is "none."

B) Find the chance of "none" (same as not A and not B and not C) using the multiplication rule.

C) Subtract that result from 1 (or 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 Expected Value and Standard Error for a Sample Percentage

With a simple random sample, the exected value for the sample percentage equals the population 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 $\frac{50\%}{\sqrt{\text{number of draws}}}$, 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.

(With a large sample, the normal curve may be used to get chances for the sample percentage.)

The Square Root Law

Multiplying the size of the smaple by some factor divides the SE for a percentage not by the whole factor—but by its square root.

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 EV and SE for the Average of the Draws

When drawing at random from a box:

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

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

(With a large sample, the normal curve may be used to get chances for the sample average.)

Using the Normal Curve to get Chances for the Average of a Large Number of Draws

When drawing at random from a box, the probability histogram for the averge of the draws follows 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.

The formula for the standard units for a given average is:

$$\frac{\text{observed average} - \text{EV for the sample averge}}{\text{standard error for the sample average}}.$$

Do not make the common mistake of using the SD instead of the SE in the denominator. True, the average of the original box and the expected value for sample average are equal, but their deviations are not.

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:

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

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.

The Gauss Model for Measurement Error

The Gauss model applies to repeated measurements on some quantity. According to the model, each time a measurement is made, a ticket is drawn at random with replacement from the error box. The number on the ticket is the chance error. It is added to the exact value to give the actual measurement. The average of the error box is equal to 0.

Since the average of the error box is 0 and each measurement is the exact value (a constant) added to the ticket from the error box, the population average weighing equals the exact value.

When two quantities differ by a constant value, they have the same SD. So the SD of the error box is the same as the SD of the measurements. When the Gauss model applies, the SD of a series of repeated measurements can be used to estimate the SD of the error box. The estimate is good when there are enough measurements.

Now, the sample average of the measurements will be off from the population average (the exact value) by the standard error for the sample average. This standard error is found by dividing the estimated SD of the error box by the square root of the 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.

A Closer Look at Tests of Significance

Data Snooping

Data-snooping makes P-values hard to interpret.

When running a test of hypotheses, first decide on a hypothesis to test and a model proposed for the population. Then choose the sample at random. If the result is consistent with the box model, the null hypothesis is accepted. Even if the null hypothesis is true, there is some possibility that the result will not agree with the box model. But that could be due to chance error. We evaluate the chance error for results that are that extreme or more extreme and decide if it is a reasonable chance error.

There is always chance error at play. We take that in mind. However, when data mining, one gets the results of various samples and picks the ones he likes to "run" the test. This is not valid chance error; the result is already known. Selecting the test to run, after the results are known is just fitting the data to a hypothesis. There was no test. The sample may have been drawn randomly, but our selecting of that test result was not random at all.

Even if the null is true, some tests will produce extreme values because of chance error. Singling in on those extreme results does not give a reason to reject the null hypothesis. If $P = 5\%$, then 5% of the results should be that extreme. It tells us nothing. Other tests that produce reasonable results should play a part in the decision.