# Statistics and Epidemiology Supplement

Author: Christian McEvoy, MD, MPH
LT, MC, USN
PGY-3, General Surgery, NMCP
mcevoycs@gmail.com
@mcevoymdmph

The following is meant to be quick reference guide on targeted topics. I hope you find it helpful. It is meant to be a supplement to the information in our various texts' chapter.

**Types of Data**

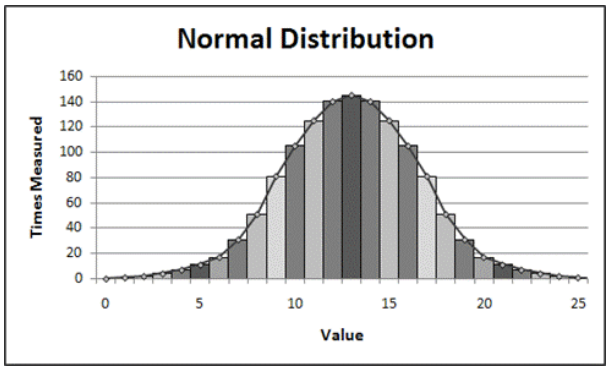
*Categorical/Nominal* → *Ordinal* → *Interval* → *Continuous*

This is really a continuum, and on standardized tests these are gimme questions if you know these definitions, but easily-lost points if you don't recall the definitions. Categorical data are nominal data. Same thing. These are data that have 2 or more categories but no specific ordering. A common example is hair color. It is nonsensical to put hair color in any order. Of note, binomial data (yes/no) are considered categorical data. Ordinal data are like categorical data in that they are categorical, but ordinal data also have an intrinsic order. For example, education or socioeconomic status (low, medium, and high). Interval data are like ordinal data except there are defined intervals between each categories. Continuing with education as an example, grades in elementary school are an example. There is 1 year between each grade (1st, 2nd, 3rd, etc). Finally continuous, variables are non-categorical. They can take on any value and are ordered. Continuous data are only limited by the measurement tool you use. For example, weight or height or age are all continuous variables.

Ratio data are a special type of continuous of interval data that have a real meaningful lowest possible value – usually zero. This makes comparisons particularly meaningful. Weight is always the classically-used example. Zero is the lowest weight. Someone who is 100 lbs is twice as heavy as someone who is 50 lbs. Try that with temperature – it doesn't work. There is no floor, so it is not meaningful to say it is twice as hot.

**Normality**

We focus on normality because *some* of our common statistical tests produce estimates and use some mathematical assumptions. One of those assumptions that some common tests use is that the data are normally distributed. When we say that our data are normally distributed, we all picture a histogram like this one →




And that is about 1/3 correct... When a statistician tests a specific variable for normality, she/he is usually performing some statistical evaluations that ask the question: *Given the data*

*I have in my data set, may I assume that this variable's values are normally distributed in the larger population?* Those tests are not asking if the data we have in a given data set fit nicely into a little histogram like the one above. This is a small, but important distinction. A statistician is not asking: are my data normal? The statistician is asking: *May I assume the data are normally distributed in the normal population?* There are many tests for this and you are unlikely to be tested on them, but they include the Shapiro-Wilk test, skewness and kurtosis testing, making q-norm plots, and even the popular eye-ball test. One special instance in which we may not safely assume normality is with small sample sizes – usually this means samples less than 30.

When we can assume normality, we use the *mean* as a measure of central tendency and *standard deviation* to describe the variation. Assuming normality, we employ parametric tests including t-tests and ANOVA and their variations. When we can't assume normality, we use the *median*, because it is not as sensitive to large outliers, and the inter-quartile ranges as well as non-parametric tests such as the Wilcoxon and Kruskal-Wallis.

**Skewness**

A close cousin to the topic of normality is the skewness of the distribution of our data. In this case, we are concerned what our data look like in graphic form. If you are anything like me, I can never remember which is negative and which is positive and which is right and which is left. And I could never remember in which case the mean > median > mode? Well there is a really nifty trick to remember this and it has only two parts.

1) Look at the long tail. Whatever side it is on, is what the skewness is named. Long tail towards the positive of the x axis or the right…then it is positive / right skewed. Long tail towards the left or negative of the x axis…then you have negative / left skewed data.
2) And the mean, median, and mode are always in the alphabetical order running away from that tail. See the graphic below.

**Comparison Tests**
The following is an abbreviated matrix highlighting the most common statistical comparison tools tested on standardized tests. This is not, in any way, meant to be comprehensive. Below the table, please find a few words on the most common tests and a few of their variations.

| | | **Central Tendency (Variability)** | **Comparing 2 groups** | **Comparing >2 groups** |
|---|---|---|---|---|
| Interval or Continuous Data | Normal | Mean (SD) | t-test or paired t-test (see notes below) | ANOVA |
| | Not Normal (non-parametric) | Median (IQR) | Wilcoxon-Mann-Whitney | Kruskal-Wallis |
| | | | | |
| Categorical or Ordinal or Binary | | Proportion | | Chi2 |

*A note about "statistical significance"
Commonly, we use p-values or confidence intervals to determine whether our results are statistically significant. It is important to remember a couple of things. These are both arbitrary in large part. As you know, convention is to set the p-value for significance at 0.05 which corresponds to probability of making a Type I error of 5% (Type I and II errors discussed later). But it is important to understand that those levels can be set by the researcher. Conventionally, we set the probability of making a Type I error at 5% and this is involved in both the p-value equation and the equation for confidence intervals. Bottom line, just remember that statistical significance is not some magic number – it is a convention but subject to change at the investigators discretion. I use the conventions in the following discussions.

**T-Test (normality assumed; interval or continuous variable)**
The t-test is used to compare means. A p-value less than 0.05 suggests that we can reject the null hypothesis that the means are equal to each other and say that they are statistically different. There is one important variation on the t-test. While the regular ole t-test is used when we are comparing the means of two groups that are not related (e.g., mean absite scores of PGY-1's vs PGY-2's), a *PAIRED t-test* is used when we are comparing means measured on the same group at different times (e.g., mean absite scores of the current PGY-5 class comparing their PGY-3 exam to their PGY-4 exam).

**ANOVA aka Analysis of Variance (normality assumed; interval or continuous)**
ANOVA is important because it is like a t-test in that it compares means of groups, but it does so across more than 2 groups. In other words, an ANOVA of just two groups is called a t-test. ANOVA can be one-way or two-way. A one-way ANOVA would compare the mean Absite scores of PGY-1, PGY-2, and PGY-3. A p-value < 0.05 would suggest that the means across all three groups are not equal to each other, but importantly, it does not point us to which group (or groups) are different. We have to test them against each other directly using a t-test to determine that. If we wanted to consider both PGY status and gender, we would use a two-way ANOVA. The two-way option allows us to use not only the PGY-1 vs. PGY-2 vs. PGY-3 status, but also allows us to look at those groups combined with male vs. female. It is two-way because we are examining two factors. Again, a p-value < 0.05 would suggest that that mean test scores across all these groups are not equal to each other, but it does not help us locate where the differences exist. There are many other variations on this type of variance analysis, and if you are interested, you can read more about ANCOVA, MANOVA. I created a short table below because after doing some test questions, I realized some of the variations of ANOVA are tested. The bolded text highlights the unique characteristics of each variation.

| Remember ANOVA used only when outcome (compared) variable is considered *Normal* | | | | |
|---|---|---|---|---|
| **Predictor(s)** | **What's Being Compared** | **Estimate Statistic** | **Test** | **Example:** |
| 1 variable w/ multiple categories | A single continuous or interval variable | Mean | One-way ANOVA (or just ANOVA for short) | Average height of student in 3 classrooms. |
| **1 group** | **A single measurement repeated more than twice** | Mean | Repeated Measures ANOVA | Average height of students in 1 classroom measured each month for several months. |
| **2 variables** each with multiple categories | A single continuous or interval variable | Mean | Two-way ANOVA | Average height of students in 3 different classrooms **but also accounting for gender**. |
| **2 variables, one with multiple categories and one continuous** | A single continuous or interval variable | Mean | ANCOVA | Average height of students in 3 classrooms **but also considering age as a predictor** |
| 1 variable w/ multiple categories | **2 or more continuous or interval variables** | Mean | MANOVA | Average height **and weight** of student in 3 classrooms |

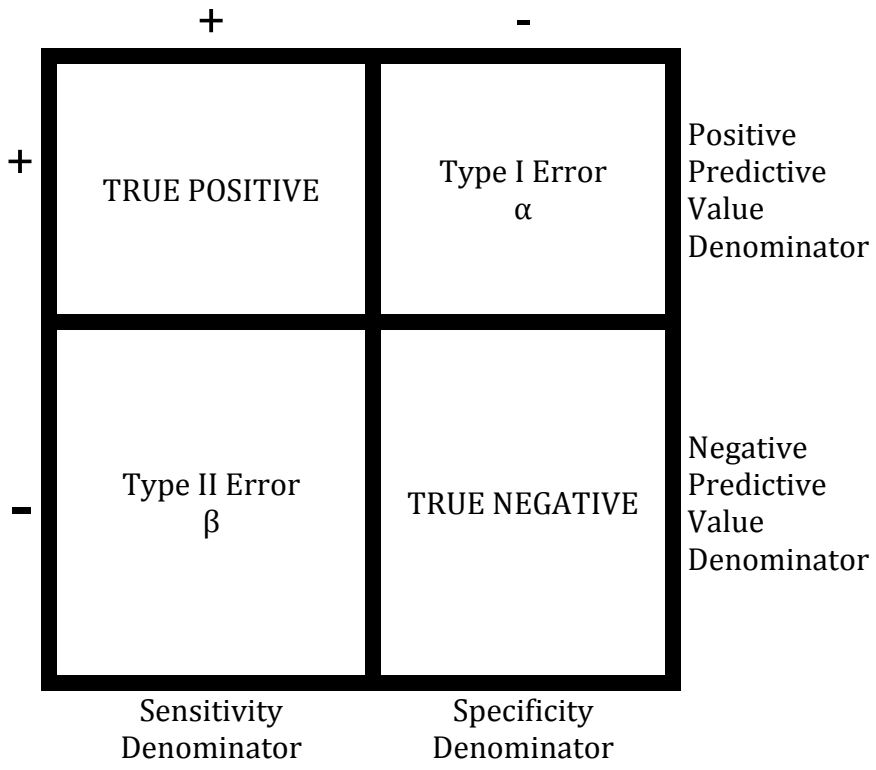
**Chi-squared (Chi2) Test (Categorical or Binary data)**
This test is used when we are comparing estimates that are expressed as a proportion. In this case, consider the Absite passing scores compared between surgical residency programs. Say NMCP has 90% passing scores for all residents and interns in the past five years. In other words, 90% of the program achieve at least a passing score. EVMS has 88%. NMSD has 86%. Note I gave percentage of passing scores - not average score, the latter would require a t-test or ANOVA. In the case of percentage of folks who passed, we can use a chi2 test to see if these values are actually statistically significantly different. A p-value < 0.05 suggests that there is a difference between these groups; however, similar to ANOVA, we don't know where. We have to test the groups against each other directly, using chi2 to determine where the differences exist. Many different named tests use a chi2 statistic. This can be confusing, but it suffices to say that these are just variations of the basic chi2. But on those lines, the following are a couple of buzzword examples to memorize for testing purposes:

-If you were conducting a *case-control study* in which you found cases of disease and matched healthy controls and compared those cases with their matched controls with respect to some binary or categorical variables, you would use a version of the chi2 test called *McNemar's Test*. You could also use this test when you have these types of variables and repeated measurements in the same subjects.

-If you had only very few observations, usually less than 30, but you wanted to perform chi2 testing, you would use *Fisher's Exact Test.*

**Tests, 2x2 Tables**

On an exam, the most common reason to encounter the following is regarding a screening test. Whenever encountering a problem involving a 2x2 table or associated calculations, regardless of the simplicity of the problem, I suggest drawing the 2x2 table complete with the margins. For calculating the Sensitivity, Specificity, Negative Predictive Value, Positive Predictive Value, remember the box over the margin rule. The box (numerator) is always the respective true positive or true negative box. The margin (denominator) is the marginal value you are calculating.




**Prevalence:** Number of cases in a given population at a defined moment in time.
= cases/population

**Null hypothesis:** (1) there is no difference between two estimates or
(2) there is no disease

**Properties of the Test**

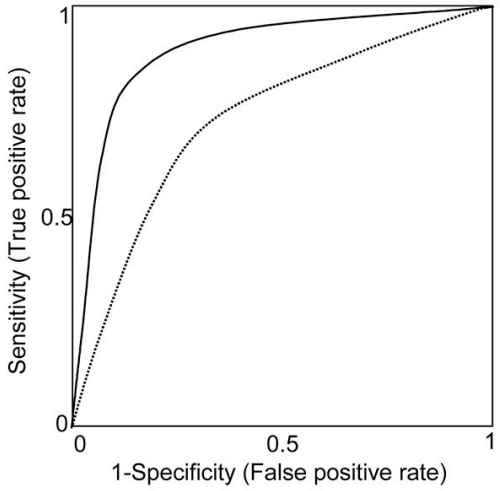
**These properties **are NOT subject** to changes in disease prevalence*

**-Sensitivity:** the ability of a test to correctly identify true positives
=true positives/sensitivity margin

**-Specificity:** the ability of a test to correctly identify true negatives
=true negatives/specificity margin

*Receiver Operating Characteristic (ROC) Curve*

I put this on here because I came across one multiple-choice question that provided this curve and asked to estimate specificity of the test. Just take a look at the x-axis and be sure you realize that it is 1 – specificity.




**-Type I error:** rejecting the null when null is true - False Positive. This error "sees" disease when there is no disease. We use **α**, the probability of making a type I error, to calculate statistical significance because, practically, if we see a difference between two groups, we want to be relatively confident that we are not seeing differences that are not there.

**-Type II error:** accepting null when null is false - False negative. This error "fails to see" disease when there is in fact disease present. We use **β**, the probability of making a type II error, to help calculate Power (1-β). Power is a test's ability to "see" a difference when it exists. Low power (or high β) means that a test has trouble detecting small differences.

**Properties of the Test *AND* Disease Dynamics**
**These properties **are indeed subject** to changes in the disease prevalence. Positive predictive value increases when prevalence increases.*

> **-Positive predictive value:** of the positive test results, how many are true positive?
> =true positive/all positive tests
>
> **-Negative predictive value:** of the negative test results, how many are true negative?
> =true negative/all negative tests

---

**Risk/Incidence**
The most direct introduction to this topic is the reminder that risk can't be calculated without time in the denominator, and in broad terms, you can't comment on causation without risk. There is obviously some nuance to this, but this is not important for examination purposes.

**Incidence:** is a **rate** and conveys risk. It is the occurrence of new cases (numerator) divided by the number of people at risk multiplied by time at risk (denominator).

*Incidence = Risk = new cases / (people at risk*time at risk)*

**When to use:** studies in which we follow subjects over time (e.g., RCTs, retrospective observational analyses, prospective observational analyses)

It bears repeating: risk is not something you can convey without time in the denominator. It is easy to slip into a discussing risk when using prevalence as the statistical estimate, and this is wrong. *This concept is testable and tested.*

**Relative Risk (RR)** = Incidence Exposed/ Incidence Non-exposed = **Risk Ratio**
> *****_Compares the risk experienced by two groups._

**Absolute Risk Reduction (ARR) = Attributable Risk =** Incidence Exp – Incidence non-exp
> *****_How much could I reduce the risk of this disease if I eliminated this exposure?_
> _*Or if the outcome is "good," how much of the observed improvement is due to the intervention?_

**Attributable Risk % =** (RR-1) / RR
> *_How much of the overall risk of this disease is due to this exposure?_

**Number Needed to Trea**t = 1/ARR **Number Needed to Harm** = 1/ARR

**NNT and NNH are equations that follow the "everything is relative rule." Let's say you have an ARR of 0.09 when comparing headache relief in people who take novel drug A vs. people who take placebo.*

*1/0.09 = 11.11= NNT*

*This NNT value indicates that we need to treat 11 people with novel drug A in order to prevent 1 headache – the key is we are looking at a treatment and a good outcome.*

*Now consider a different example. Suppose you have an ARR of 0.09 when comparing people who smoke to people who do not smoke, and the outcome of interest is headache (note not headache relief). In this case, we are examining a risk factor and a bad outcome. We use NNH, but get the same point estimate.*

*1/0.09 = 11.11= NNH*

---

**Prevalence/Odds Ratios**

**Prevalence:** cases currently with disease / total population

*Note time is not in the denominator, so risk is not evaluated

**When to use:** studies in which we get snapshots of time but do not follow over time (e.g., case-control studies, cross-sectional analyses)

**Odds:** ratio of cases to non-cases = cases/non-cases

*If you have 100 people and 2 have the disease, the odds are 2/98

**Odds Ratio:** A ratio of ratios = odds exposed / odds unexposed

**This is used to evaluate the odds (not the risk) of being a case when exposed to certain risk factor.*

**Odds Ratios vs. Relative Risk Ratios**

| What is the same? | What is different? |
|---|---|
| *-They are both ratios comparing exposed groups to unexposed groups.*<br><br>*-Both have their statistical bona fides characterized by confidence intervals. A confidence interval that does not include 1 indicates "statistical significance in both."* | *- RR compares risk (includes time in the calculation) while OR compares prevalence. RR suggests may suggest causation. OR may only suggest correlation (except in one circumstance)*<br><br>***Denominators of Odds and Risk***<br>Odds denominator = non-cases<br><br>Risk denominator = (cases + non-cases) x time |

*Note: there is one instance in which the OR may by used to estimate the relative risk - when the disease in question is very very rare. The usual rule is < 10% prevalence. The following example shows the math:

*Rare disease with < 10% overall prevalence*

*Let's compare:*

*Group A (exposed) 2% prevalence of disease in group A*

*Group B (non-exposed) 1% prevalence of disease in group B*

*Odds exposed in A = 2/98 ; Odds non-exposed in B = 1/99*

*Odds Ratio = (2/98) / (1/99) = .0204 / .0101 = 2.10*
*= 2.1x odds of rare disease found in A vs B*

*Now let's pretend we were able to follow the same two groups for 3 years and we found the risk was similarly 0.02 and 0.01 respectively.*

*Risk Ratio = (2/100*1) / (1/100*1) = 2.0x risk in A compared to B*

*OR 2.1 ≈ RR 2.0*

*This approximation works because the risk denominator is so small in comparison to the numerator. If you try the math for a prevalence of greater than ~10%, you will see it does not work out so well.*

**Standard Error vs Standard Deviation**

Sometimes it is easy to confuse standard error (SE) and standard deviation (SD), and I've seen questions that try to trip you up on this difference. SD expresses *variability* in your data. In other words, SD looks at the mean and tells you how far your data stray from that mean. SE is also known as the standard deviation of the mean. It is *not a measure of variability*, but rather an estimate of how accurate the estimated sample mean is when compared to the real population mean. If our PGY-1 class took the Absite and the average score in our class was 55%, getting a standard deviation would tell us if we were all close to that 55% or if we had high and low scores that just balanced out on average to 55%. A large standard deviation would indicate the latter. A standard error estimate, on the other hand, would tell us how likely it is that our mean score reflects the mean score of all other PGY-1s in the country. The standard error is really a function of the sample size. As sample size increases, it displays more confidence that our sample estimate is closer to the population reality. So if we included all US Navy PGY-1 Absite scores in our mean score, our standard error would decrease.

**Standard Deviation (σ) =** $= \sqrt{\frac{\sum(x - \bar{x})^2}{n - 1}}$

**Standard Error** = $\sigma / \sqrt{n}$

**95% Confidence Interval**

$$\bar{X} \pm 1.96\, \sigma/\sqrt{n}$$

*This equation states that we calculate the C.I. for a mean by using the standard deviation and the sq root of the sample size. 1.96 is the scored translation of a two-tailed alpha level (probability of Type I error) of 5%. The important thing to notice here is that our confidence interval gets much smaller (we get more confident) when same size increases.