Chapter 2: Tools, Tool Use, and Associated Pain

The major goal and objective of this chapter is to review common statistical tools needed to link HR practices to business outcomes with as little pain as possible. Ok, about now is when most people start to get nervous. I can’t promise it will be “pain free,” if only because of the negative emotional baggage you brought with you from prior exposures! So, I want to first touch on some all too common experiences and consequent feelings many business professionals have after their initial – generally unpleasant - exposure to statistics. I am going to assume most readers have completed (or are in the process of completing) the equivalent of the core curriculum requirements in quantitative methods found in accredited U.S. business schools. Unfortunately, in my experience most individuals emerging from these courses did not enjoy their experience and cringe at the mention of “null hypotheses” – this is partially why the word “statistics” is not in the title of this chapter! The number of reasons contributing to this experience would fill a chapter by themselves. A short list of influences making initial statistics exposure so bad includes:

1. instructors who are so impressed with themselves that all their explanations use a half a dozen 5 syllable words (e.g., how often do you hear “heteroskedasticity” in a sentence every day? . . . how many of you knew there are two correct spellings, the other of which swaps a “c” for the “k”?),

2. uninspired teaching assistants, or worse, uninspired teaching assistants who speak English poorly as a second language (is it my imagination, or did most of these folks take subsequent jobs with call centers?),

Chapter 2 Objective: to review two common statistical tools needed to link HR practices to business outcomes with as little pain as possible. Statistics reviewed include those sensitive to:

- Whether differences exist.
- Whether relationships exist.
3. use of uninteresting and uninspired examples (I get nauseous at the mention of urns containing red die and white die); and/or,

4. mindless, repetitious drill of weekly problem sets without proper care taken to establish the pattern recognition needed to set up problems correctly.¹

I could retire if I had a nickel for every time a student said “if you just told me how to set the problem up, I can apply the formula correctly and come to the right answer.” Unfortunately, the answer to this is “if I set it up, why do I need you? I have a computer that can do it better than you can – the whole point is for you to learn how to identify which questions need to be asked of the data and what statistic(s) gives an answer to those questions.”

The remainder of this chapter is about statistics and research design in applied HR settings. I took my first statistics course in the summer of 1974 at the University of Iowa from Dr. James Maxey at 7:00 am Monday through Friday for eight weeks. Dr. Maxey was the first of many professors at Iowa who captured my interest in statistics by introducing real problems and issues first, then showing how statistics can provide insight and/or solutions to those problems. I owe an enormous debt to Drs. Maxey, Feldt, Foresyth, Novick, Cleary, and many others who provided this delightful foundation in applied statistics. My goal for the remainder of

¹ Gigerenzer’s (2004) discussion of the “null ritual” is the logical outcome of these mindless repetitions. My point, and that of Gigerenzer, Fisher (1955, 1956), and many others is that application of statistical convention absent appreciation of existing knowledge of the phenomena of interest is foolish and detrimental to understanding.
this chapter is to approximate for you what my faculty at Iowa did for me. Hence, using a realtor training example to kick things off below, I start with HR-BOut inferences of interest and work backwards toward associated statistical and research design issues. Note, the BOut of interest will solely depend on what is of most strategic value to the organization in question. BOut measures are only limited by the imaginations of HR and line management and willingness to exert the effort needed to actually obtain error free assessments of those measures. BOut can include, but are certainly not limited to, number of units produced, scrape rate, inventory costs, voluntary turnover rate, 30/60/90/ day job survival, expected job tenure, market penetration, profit, sales, etc. In other words, BOut constitute any and all economic or operational measures relevant to the firm’s strategic goals and objectives, with emphasis on the phrase “relevant to the firm’s strategic goals and objectives.”

Table 2-1 below provides examples of key HR-BOut questions and the type of statistics that might answer those questions. Note, all of the statistics involve one or more of the following:

- Differences between two or more groups.
- Correlations between a BOut measure and one or more measures obtained from an HR practice (e.g., applicants’ scores on some selection test).
<table>
<thead>
<tr>
<th>HR-BOut Question</th>
<th>BOOut of Interest (cost and value added)</th>
<th>Statistics Reflecting HR-BOOut Relationship of Interest</th>
</tr>
</thead>
<tbody>
<tr>
<td>How does our “merit pay” system relate to employees decisions to work hard and/or continue employment with the firm?</td>
<td>➢ Units produced. ➢ Scrap. ➢ Inventory costs. ➢ Voluntary turnover rates. ➢ Increased cost of merit pay system due to actual change in pay levels and any administrative overhead. ➢ Average pay variance under original and merit pay systems.</td>
<td>➢ Differences in average BOOut between divisions using vs. not using the merit pay system ($H_o : \bar{X}<em>{\text{merit pay}} \leq \bar{X}</em>{\text{no merit pay}}$ vs. $H_A : \bar{X}<em>{\text{merit pay}} &gt; \bar{X}</em>{\text{no merit pay}}$). ➢ Estimates of how each BOOut changes for each $1 allocated for merit pay (slope of the line obtained from regressing BOOut onto merit pay, or $b_1$).</td>
</tr>
<tr>
<td>How does our new “executive development” mentoring system targeting middle management work?</td>
<td>➢ Promotion rate. ➢ Annual bonuses received. ➢ Unit sales/profit/market penetration. ➢ Costs associated with new development system.</td>
<td>➢ Differences in BOOut across individuals “developed” with the new system or units employing those individuals ($H_o : \bar{X}<em>{\text{mentoring system}} \leq \bar{X}</em>{\text{old system}}$ vs. $H_A : \bar{X}<em>{\text{mentoring system}} &gt; \bar{X}</em>{\text{old system}}$). ➢ Estimates of how each BOOut changes as a result of each $1 spent on “executive development” (slope of the line obtained from regressing BOOut onto dollar cost of development, or $b_1$).</td>
</tr>
<tr>
<td>How well does our new diversity initiative work?</td>
<td>➢ % of diverse individuals targeted jobs in the firm relative to their representation in the relevant external labor market. ➢ Economic outcomes including cost savings, profit, sales, customer satisfaction, market penetration, etc. ➢ Costs incurred in diversity initiative.</td>
<td>➢ Differences in each BOOut across more vs. less diverse employee work teams ($H_o : \bar{X}<em>{\text{diverse teams}} \leq \bar{X}</em>{\text{nondiverse teams}}$ vs. $H_A : \bar{X}<em>{\text{diverse teams}} &gt; \bar{X}</em>{\text{nondiverse teams}}$). ➢ Estimates of how each BOOut changes as a result of each $1000 spent on diversity initiative (slope of the line obtained from regressing BOOut onto dollar cost of diversity effort, or $b_1$).</td>
</tr>
<tr>
<td>How well will the new call center selection system work?</td>
<td>➢ Average job tenure and recruiting/training costs associated with higher job tenure. ➢ Average customer satisfaction survey rating. ➢ Average weekly sales volume.</td>
<td>➢ Correlation ($r_{xy}$) between selection system scores and BOOut measures. ➢ Expected increase in job tenure or decrease in recruiting/training costs resulting from use of system. ➢ Expected increase in job tenure or decrease in recruiting/training costs expected with each 1 point increase in an applicant’s selection score.</td>
</tr>
</tbody>
</table>
The nature of some HR policy/practice → BOut relationship differs across groups, as captured by differences in slopes (the amount by which some BOut measure is expected to change when there is a 1 unit change in a measure X obtained from some HR policy or practice) and constants (i.e., the expected BOut level when X = 0) in estimated models of HR policy/practice → BOut relationships.

Now, this is not meant to be a comprehensive review of every statistic or probability distribution you might encounter (or need) in your tool box when looking for HR-BOut links. You have been exposed to many additional tools already in your statistics courses (e.g., F tests of whether three group averages are equal). Additional “tools” that you have already been exposed to will be reviewed as real case examples testing HR-BOut links and are examined in later chapters on recruiting, selection, training, compensation administration, etc. The remainder of this chapter will attempt to provide a “minimally painful” review of use of simple differences, correlations, and slopes (i.e., ordinary least squares linear regression analysis). I urge the reader interested in a book-length review of regression to see Philip Bobko’s Correlations and Regression text (the only statistics text I have ever used that students said they enjoyed reading!).

**Business Outcome Criteria**

We would probably want to consider the fact that things like interest rates, local unemployment levels, local economic growth, etc. will all affect housing sales and might be different for those receiving training under the old versus new training system. Discussion with management would focus on the realty firm’s strategic goals and objectives and how we would know whether a realtor contributed to them or not. The following three “performance” criteria might result from such a discussion:

1. amount by which first year sales volume differed from average annual sales volume for all “seasoned” realtors in the area for the same time period;
2. amount by which number of sales closed in the first year differed from average number of sales closed by all “seasoned” realtors in the area for the same time period; and,
3. amount by which average sale price for first year sales differed from average sale price for all “seasoned” realtors in the area for the same time period.

Assuming the old training system was discontinued when the new training system was implemented, we need performance metrics that “control” for differences in interest rates and other economic conditions that existed when each particular training program was in use by taking into account performance attained. We do this by taking into account performance by “seasoned” realtors under the same conditions – presumably interest rates, economic conditions, etc. effect newly hired realtors and seasoned realtors equally. Comparing these performance measures for newly hired realtors trained under the new and old systems will tell us which training system results in newly hired realtors coming closest to performance exhibited by “seasoned” realtors (again, assuming performance levels of seasoned realtors is desirable).
An Example of Differences: Is the new HR system better?

Perhaps the simplest way of linking HR practices to business outcomes involves examining whether a better business outcome occurs with the introduction of some “new and improved” HR policy or practice. Many questions about HR practice involve simple “differences,” as in “will our new sales training program for newly hired realtors yield different (higher) performance in the first year of employment than those trained using our old training program?” As noted above, before we actually start mucking around with any of the foul stuff you disliked so much in your statistics classes, we have to figure out what differences are important to consider, i.e., we would first have to arrive at one or more measures of “performance.” The Business Outcome Criteria sidebar discusses issues and processes in developing good performance measures, i.e., performance measures that don’t fail to capture some important component or dimension (deficient) or contain some extra stuff that is unrelated to performance (contaminated). Three possible examples of what might constitute good performance measures coming from the sidebar discussion include:

1) amount by which first year sales volume differed from average annual sales volume for all “seasoned” realtors in the area for the same time period;
2) amount by which number of first year sales closed differed from average number of annual sales closed by all “seasoned” realtors in the area for the same time period; and,
3) amount by which average sale price for first year sales differed from average sale price for all “seasoned” realtors in the area for the same time period.

Close examination of these three performance measures reveals the first two are measures of different performance quantities (dollar volume and number of sales), while the third is a measure of relative performance (average dollars per sale). Which one is more or less important will depend on the realty firm’s strategic goals and objectives. For example, young firms with a

---

2 Generically, this question could be rephrased as “does the new (training, recruiting, compensation, selection, benefits, etc.) HR system yield better business outcomes than the old HR system?”
growth strategy might initially focus on sales volume or number of sales in order to achieve some critical level of market presence, while mature firms with an already established market presence might focus on efficiency and profits by maximizing average sale price.

Ok, now that we have our measures we can make some comparisons to see if realtors trained under the new system perform better than those trained under the old system. One could simply calculate averages for these three performance metrics for, say, 50 newly hired realtors who went through the new training program and achieved at least one year of job tenure. We could then compare these performance metrics to the same measures calculated on all newly hired realtors trained under the old system over the last 3 years or so (say, the last 150 newly hired realtors who went through the old training program and subsequently achieved 1 year of job tenure). Anyone looking at these averages could immediately see which training system yielded higher performance, as the averages are not likely to be exactly equal to the penny (in many markets averages for the first and third criterion measures might be in the millions of dollars). One could choose to simply use the training system with the highest average first year performance metrics.

---

3 Why more (150) with the old system and less (50) under the new system? We could wait ~ 3 years for 150 to be trained with the new system, but that means we have to wait an extra ~2 years to get the answer to our question. If the old system works as well or better than the new system, 2 extra years of new hires would have been trained under the new system when they could have gone through the old system. I will address compromises between preferences for larger sample sizes (more is always better) and organizational realities at a number of points in latter chapters. For right now, the answer is that information on the 150 who went through the old system is already available, while we might have to wait 2 or more years to get that many under the new system.

4 Right now you are thinking “what if one training system was better on 2 of the metrics while the other system was better on the 3rd metric?” I address that issue in at least one example in all subsequent chapters. For this first example, I will keep it simple and assume all differences are in the same direction.

5 I know, now you are thinking “what if one training system costs a lot more than the other?” Remember, I want to keep this first example simple, so I will assume costs for the two training systems are equal.
I wish it was that simple. Yes, the average first year performance metrics under the new training program may be higher for this group of 50 newly hired realtors. If that was all you wanted to know, you could stop here. However, deciding to use the new training program will affect ALL future newly hired realtors, not just this particular set of 50. Recall the original question asked whether performance for newly hired realtors was higher under the new training program. The question did not ask whether performance for this group of 50 newly hired realtors was higher. The question requires that we draw an inference from performance of the first group of 50 realtors going through the new training program to all future realtors who might go through the program. Any future group of 50 newly hired realtors trained with the program will probably have slightly higher or lower average first year sales volumes when compared to this first group of 50. In fact, we might have just been lucky with this first group of 50 and gotten the best 50 of all newly hired realtors for the next 10 years! Maybe the average performance metrics for the next 10 sets of 50 newly hired realtors are actually well below the

The “pling” in Sampling Distribution
Remember “sampling distributions” in your statistics courses? I underlined the “pling” to distinguish what we are talking about from the sample distribution. We have two sample distributions in this example composed of the 50 or 150 sales volumes, number of sales, and average sales. Each can be characterized by some measure of central tendency (e.g., the mean) and spread around the mean. When we change the word “sample” to “sampling” we are no longer talking about these two samples of 50 and 150. Instead we (typically) are talking about the distribution of means we would get if we took, say, 10,000 random samples of 50, calculated the mean on each one, and looked at the distribution of these means (similarly, we could have taken 10,000 samples of 150). Of course, we don’t do that, because if we did we would not say we had 10,000 samples of size 50, instead saying we had a sample of 500,000. Regardless, our question about which training program yields higher average performance focuses on a “group” as the relevant unit of analysis, i.e., all possible future groups of 50 newly hired realtors. We are not asking whether “Jim” or “Sally” or some other individual will perform better after new vs. old training programs. Whether a group will perform better under new vs. old training systems is best answered by comparing group-level statistics, e.g., sample means. The best “group” level statistics available come from the samples of 50 and 150 we have at hand. Not surprisingly, we will estimate what the sampling distribution will look like from sample statistics a little later on.
average of this first set of 50. You certainly don’t want to use the new training system if it only yielded higher performance for the first 50 realtors who went through.

In other words, any difference you see in performance metrics could be due to either 1) true differences in average performance caused by the training programs (i.e., the answer you hope to find to the question asked) or 2) differences in average performance metrics one would expect to occur by random chance across multiple samples of 50 trained under the new system and multiple samples of 150 trained under the old system when the old system works as well on average as the new system. Random chance could have caused particularly high ability realtors to have been among the first 50 to go through the new training system. Alternatively, random chance could have caused the last 150 to go through the old system to have been particularly low ability realtors (see the sidebar discussion of sampling distributions). If either of these circumstances occurred, random chance will have caused the samples of 50 or 150 to not be representative of the larger population of new realtors. Since we again want to draw an inference about whether training differences will occur in ALL groups of new realtors, we need some way of determining whether a difference we see between these groups of 50 and 150 are due to random chance or due to true differences.

So, what you really want to know is whether the performance metrics gathered from the first 50 newly hired realtors are so much bigger than performance metrics gathered from 150 realtors trained under the old system that they are unlikely to have happened by random chance, i.e., they are likely to have occurred because of true performance differences between the training programs. As one might suspect, small performance differences are more likely to have

---

6 Of course, any difference could be due to the effect of some influence other than the training program. One more time – recall I am trying to keep this example simple, so I will, again, hold discussion of this kind of complexity for latter chapters.
occurred by random chance, while large performance differences are more likely to have occurred due to true performance differences caused by the new and old training programs. It might help to think of this as a kind of “backwards” logic. Specifically, to convincingly end up with the conclusion you hope to draw (i.e., that the new training system generated better performance), you start with the conservative initial position that there is no performance difference. However, after seeing how large the difference is in average performance metrics is between the samples obtained using the new and old training systems, you are so overwhelmed by evidence showing the performance difference is not 0 that you drop your initial, conservative position and adopt a new position, i.e., that there is a “significant” difference and the new training system is better. Instead you conclude that the average performance metric is so much bigger for the 50 receiving the new training system that it is unlikely to have happened by random chance when true performance is actually equal across populations receiving the two types of training.

**Statistical Significance:** If the difference is large enough that it is unlikely to have occurred by random chance, we say it is “statistically significant.” Unfortunately, with large enough samples, even the smallest difference can become statistically significant. Of equal, if not greater interest, is whether the difference is large enough to be “meaningful.” More on this latter.
Again, in this instance the question of interest applies to the population of all future groups of realtors who might be newly hired into the firm and trained with one of these two systems. However, we don’t have information on how all future realtors who might be trained under either system perform – we do have information on how 50 newly hired realtors performed after receiving the new training program and how the last 150 newly hired realtors trained under the old system performed. Average performance metrics for the sample of 50 who went through the new training system are our best estimates of how the population of all groups of 50 newly hired realtors might perform in the future if they go through the new training program. Average performance metrics for the last 150 who went through the old training system are our best estimate of how the population of all groups of 150 newly hired realtors might perform in the future if they go through the old training system.

Null Hypothesis About a Difference. Believe it or not, the paragraph above coarsely describes the logic and process associated with a one-tailed test of the null hypothesis

\[ H_0 : \bar{X}_{new} \geq \bar{X}_{old} \] versus the alternative hypothesis

\[ H_A : \bar{X}_{new} < \bar{X}_{old} \] (remember, \( \bar{X} \) is the average difference between performance of newly hired and seasoned realtors . . . we expect this difference to be smaller among those receiving the new training because those newly hired

**Why not use everyone?** One might consider using performance metrics available on ALL newly hired realtors who completed the old training program and at least one year of subsequent job tenure. However, this might involve \( N = 1,000 \) individuals hired and trained over the last 25 years. While more information is clearly preferable to less information, the further back in time we go to gather “old training system” performance data, the more likely it is that the data will be contaminated by some source of influence or effect that is no longer present, i.e., that will NOT affect those trained under the new system. For example, using performance data for those trained under the old system obtained since 1987 would include the effects of the “black Monday” stock market crash in October, 1987. Assume for the moment that seasoned realtors know things to do to survive severe medium term market downturns that novice realtors don’t know. If performance data obtained on new realtors trained in the fall of 1987 is included for comparison purposes, average seasoned-novice performance differences are likely to be larger than they otherwise would have been, causing the new training system to look better (i.e., yield smaller seasoned-novice differences) by comparison. Ultimately it becomes a judgment call trading off the increased sample size obtained by going farther back in time to obtain performance data on the old system versus the risk of inadvertently increasing contamination (error) in your performance measure.
realtors will perform at levels closer to that achieved by the seasoned realtors). Right now I expect most readers are experiencing a mild shudder, recalling painful memories from required statistics courses . . . things like Greek letters (e.g., Σ, σ, μ, α, β, and ρ), the labels “null” versus “alternative” hypotheses, weekly required homework assignments, z statistics and looking up probabilities using the normal probability density table in the back of their textbook, etc. I will finish the realtor training example below, describing the steps already covered using formal statistical descriptions you all saw in earlier statistics courses, as well as describing the statistic used. I will then briefly outline the specific tools and procedures to be reviewed in remainder of this chapter, which is limited in focus to three fundamental statistics used to answer questions about 1) differences and 2) relationships. I discuss some more advanced tools in the appendix for those who might be interested. “Advanced” tools are used in the last and most complex case example in each chapter.

**Bringing Home the “Differences” Example.** Ok, the realtor training example focused on whether a new realtor training program yielded better performance outcomes than an old realtor training program. Any actual differences observed could be due to 1) true performance differences resulting from the two training programs or 2) random chance. We want to know if the 50 realtors completing the new training system performed so much better than the last 150 realtors trained under the old system that it couldn’t have happened by random chance alone and had to be due to the new training system. Well, the bad news is we can never know that for certain unless we somehow obtain information on first year realtor performance of every possible newly hired realtor who might be hired in the future (i.e., unless we get information on the whole population of interest). Yes, there will always be a small possibility that any difference, no matter how large, might have been due to random chance effects that caused the
samples not to be representative of the population(s) they were drawn from. Over 30 years of experience linking HR practice to business outcomes has shown me that some pretty strange things can happen by random chance. Remember to be humble . . . a statistically significant finding does not guarantee true insight into the phenomena of interest! It is but one, very limited, view of the phenomena of interest, and must be considered in light of all other information available. Again, more on this latter.

A different way to ask the question is “given the new training system yields higher performance outcomes that are very close to that achieved by seasoned realtors, how much closer to the performance level achieved by seasoned realtors do they have to be before we are comfortable enough to assume the new training system will work better in all future newly hired realtors?”, or “how much risk of being wrong are we willing to take, i.e., risk that our conclusion is wrong and the new training system does not yield higher performance?” For example, assume the new training program yields an average annual sales volume that is $1.75M closer to seasoned realtor sales volume than the old system. If I say there is a 1% chance that random samples of 50 and 150 newly trained realtors could have an average difference of $1.75M when those samples are drawn from populations that actually have no difference in performance, are we willing to take that 1% risk (i.e., risk incorrectly concluding 1% of the time that there is a true difference in sales volume and the new training system is better)? Are you comfortable enough with a 5% chance of being wrong, a 2.5% chance, at .1% chance? This is always a judgment call by the investigator, and brings home the importance of my earlier point -

---

7 See the work of Tversky and Kahneman (1986) on framing effects to see how simple changes in wording used to describe a probability affect whether it is an acceptable risk. Not surprisingly, people react differently when asked whether they prefer purchasing a drug to cure an outbreak of disease in a rural village of 1,000 people when you say “20% chance all people will die if the drug is administered” versus “80% chance all people will live if the drug is administered” (note, both are simultaneously true, though one is positively and the other negatively framed).
these judgments are not made in a vacuum with only the statistic of interest staring at up at you from a computer screen.

Some of you may recall that common traditionally “acceptable” levels of this “risk” include 5%, 1%, and .1% (often reported as “\(p < .05 \text{ or } .01 \text{ or } .001\)”). If the difference is so large that it is likely to have occurred by random chance only 5% of the time, we say this difference is “statistically significantly different from 0 at \(p < .05\).” You may also recall that this kind of mistake – rejecting our initial, conservative position of no performance difference when we shouldn’t have rejected it - is a “Type I error.” Incorrectly concluding the difference was not particularly big and was probably due to random chance when in fact there is a difference in the larger population was labeled a “Type II error.”

Types I and II error are just fancy labels for the two most obvious mistakes we could make in drawing inferences about a population from sample data. Using non-fancy language, a Type I error here involves making a proactive mistake - adopting the new realtor training system because you think the sample difference is so big that it can’t be due to random chance variations across samples, but in fact the difference was due to random chance and the new training system really isn’t any better than the old system (Type I error – incorrectly rejecting the null hypothesis). Again using nonfancy language, a Type II error involves what economists call an opportunity loss – because we thought the sample difference was too low (i.e., non-significant), we fail to take advantage of the opportunity offered by the new realtor training program when in fact the new training system would have been better. (Type II error – failing to reject the null hypothesis when you should have).

z Statistic Testing Whether a Difference Exists

---

8 Note, there are Type III and Type IV errors too – see XXXX and ZZZZ for a detailed description of controversies surrounding the use of tests of statistical significance.
Now, let’s look at the actual statistic used to determine whether the difference is “statistically significant.” One useful characteristic of the statistic in Equation 1 is that it gets bigger as the difference in training performance gets bigger – this can be quickly seen from the fact that the numerator (the part on top of the biggest division sign) is simply the difference in average performance for 50 receiving the new and 150 receiving the old training system.

\[
z = \frac{x_{new} - x_{old}}{\sqrt{\frac{\sum_{i=1}^{50} (x_i - \bar{x}_{new})^2}{50 - 1} + \frac{\sum_{i=1}^{150} (x_i - \bar{x}_{old})^2}{150 - 1}}}
\]

Equation 1

However, the stuff in the bottom of Equation 1 looks evil! Again, let’s break it down. Look first at the stuff inside the curved parentheses and behind the Σ, i.e., \((x_i - \bar{x}_{new})^2\) and \((x_i - \bar{x}_{old})^2\). Each of these describes squaring the difference between each individual’s performance (\(x_i\) represents the performance outcome for person \(i\), where \(i\) ranges from 1 to 50 in the sample receiving the new training and from 1 to 150 in the sample that received the old training) and the average for the sample that individual came from. The symbols \(\sum_{i=1}^{50}\) and \(\sum_{i=1}^{150}\) simply ask you to add up (or sum, hence use of the Greek capital letter S) the squared differences from the mean from each sample (remember the label “sum of squares?” – this is where it comes from). The more spread or dispersion in a sample around the mean, the larger the sum of squared differences will be. The sum of squared differences around the mean is then divided by the number of squared differences that had been added up (minus 1)\(^9\) in each sample, which

\(^9\) The reason each sum of squares is divided by the number of squared differences that had been added together minus 1 is because \(N - 1\) describes the amount of independent pieces of information that went into each sum. It
Figure 1 gives you something like an average of the squared distance each person’s performance was from the mean within each sample. You may recall this was called sample “variance” in your prior statistics classes (often written as $s^2$ when calculated within a sample). The two variances, weighted by their respective sample sizes (the 50 and 150 at the very bottom of Equation 1), are added together in what is often called a “pooled” estimate of variability. Finally, the square root of this pooled estimate of average squared differences from the sample means is taken.  

So, what is this monstrosity in the denominator and why is it there? Looking at its decomposed parts suggests it somehow captures how individual performance varied within each sample. If individual performance is spread widely around the mean in each sample, the square root of these average sums of squared differences will be large. If individual performance is tightly packed right around the mean, the square root of these average sums of squared differences will be small. Why do we care about each sample’s variability around the mean? Figures 1 portrays situations in which the old and new training programs have the same difference in

---

10 Why take the square root? Recall the bottom part of Equation 1 is supposed to be a measure of variability of the simple difference of interest in the top part of Equation 1. Since we squared and summed the differences between each individual’s performance and the average performance for his/her group, taking the square root puts the denominator back into the original measurement scale. Judging whether the difference between sample sales means is due to random chance effects of sampling error requires comparison of that difference to a measure of sales mean variability, not average mean sales variability squared! Talking about variability within the sample in terms of squared average sales volume, squared average number of sales, or squared average sales volume per sale makes no sense. Comparing the actual difference in average sales volume, etc. to its variability measured on the same scale makes sense.
average performance, but differ in variability. Which are you more confident suggests that a true difference exists in the population? Hopefully you chose the blue frequency distribution in Figure 1. While both blue and red distributions show the same average difference between the new and old training systems, the blue distribution shows a situation in which variability due to random chance is less likely. In other words, it shows a situation where the mean sample difference between the old and new training systems is more likely to have been caused by a true difference in mean performance in the populations the samples were drawn from (i.e., all possible future newly hired realtors who go through the new versus old training programs). The red distribution shows a situation where the mean difference between old and new training systems is more likely to have been caused by random chance.

The denominator of Equation 1 is often called the standard error of estimate simply because the bigger it is, the more likely the sample average is not a good estimate of the population average (and, hence, is in “error”). Since this formula “pooled” estimates of variability from two samples, it is strictly speaking a standard error of estimate for the difference between two sample means created by “pooling” standard errors from each individual sample’s mean (note, either portion of the denominator by itself is the standard error of estimate for each individual sample’s mean).

So, why create a z statistic dividing the difference by its standard error of estimate? How does this help us figure out how likely it is that a given difference would have occurred by random chance when the two population averages are really equal? The answer lies in something called the Central Limit Theorem – it tells us pretty close to exactly what the probability is that a z statistic we calculate on a sample will fall in any given range of values (e.g., be greater than z = 2.0). Hopefully, many of you recall the “normal” curve. Unfortunately,
there are few things in nature that actually look “normal” when you plot their frequency of occurrence. The Central Limit
Theorem is a proof showing that something made by man – the z statistic in Equation 1 – occurs with very close to the same probability as that found in the normal distribution when sample size is “large,” typically greater than 30. Or, as noted earlier in this chapter, the sampling distribution, i.e., a theoretical distribution of sample statistics obtained from an infinite number of samples of size N, quickly approaches the normal distribution when N = 30 or greater. Luckily, we have 50 and 150 realtors in our two training samples, causing z in Equation 1 to be very, very close to normally distributed.11

What do you do with “small” samples? As noted in the text, the normal distribution cannot be used to draw inferences about mean differences with the sample size is “small,” or typically when $N_1 + N_2 < 30$. Unfortunately, this is often the case when examining HR→BOut linkages. A small realty firm may only have data on 19 newly hired realtors who went through the old training system and 10 who went through the new training system. In this instance, a different statistic can be used if one assumes the populations the samples were drawn from are normally distributed and the two populations the samples were drawn from have equal variances. Of the two, the assumption of equal variances is more important, and can be tested before hand (see the appendix for an explanation of the test). If these assumptions can be reasonably made, the test statistic of interest becomes:

$$t_{N_1+N_2-2} = \frac{\bar{x}_{new} - \bar{x}_{old}}{\sqrt{\frac{N_1}{\sum (x_{i,new} - \bar{x}_{new})^2/(N_1-1)} + \frac{N_2}{\sum (x_{i,new} - \bar{x}_{new})^2/(N_2-1)}}}$$

Now, the explanation in the text for the Z statistic basically holds here too – the stuff on top is the same $(\bar{x}_{new} - \bar{x}_{old})$, while the ugly stuff on the bottom again reflects the variability shown in Figures 1a and 1b. Notice this t statistic has a subscript $N_1 + N_2 - 2$, which is referred to as its degrees of freedom. Instead of using the normal distribution in the back of your statistics textbook, you turn to a table for t-distribution probabilities. t-distribution probabilities are typically shown just for usual $p < .05$, .025, .01, and .001 levels, because there are different probabilities depending on the degrees of freedom. In this instance we would have to find the column (or row, depending on the t-table’s layout) for df = 27, then identify the value of the t-statistic for $p < .05$ from the table ($t_{26,.05} = 1.703$). If the t statistic we calculate from our sample data is greater than 1.703, we reject the null hypothesis and conclude the new training program is yielding higher performance levels. Notice the value from the t-distribution table (1.703) is bigger than the critical value taken from the normal distribution table (1.645). As sample sizes get smaller and the t-distribution has to be used, bigger mean differences are required before one can conclude with the same confidence that any observed difference did not occur by random chance.

---

11 Interestingly, the square of a normal variable is distributed along what is called a chi-square distribution, and the ratio of two squared normal variables is distributed along what is called an F-distribution. Statistics that follow these distributions are useful in examining other kinds of “difference” questions described later.
So, after computing the z statistic from the two sets of sample observations, we look to see where this z statistic falls on the normal curve. If \( z = 0 \), then \( \bar{x}_{\text{new}} - \bar{x}_{\text{old}} = 0 \), and 50% of the normal probability distribution falls above (or below) \( z = 0 \). In other words, when the null hypothesis is true and there is no difference in average performance for those trained under the new versus old systems, we would expect to see a z statistic greater than zero 50% of the time and less than zero 50% of the time. Alternatively, if \( z = 1.96 \), we would expect to see \( z \geq 1.96 \) only 2.5% of the time when the null hypothesis is true. You may recall this step from prior statistics classes, as this is where you look up a probability associated with the z statistic in a table of normal probabilities usually reprinted in the back of your statistics textbook. For example, if \( z = 1.645 \) you immediately know two things. First, the new training system yielded higher average performance than the old training system, otherwise the z statistic would be 0 (if they were exactly equal) or negative (if the old training system yielded higher average performance). Second, you look up the value for \( z = 1.645 \) in your normal probability table and discover that 95% of the normal distribution is expected to fall below this value. In other words, if the null hypothesis is correct and the true population difference is 0 or negative, then differences as big or bigger than what we actually saw between these two samples (i.e., \( \bar{x}_{\text{new}} - \bar{x}_{\text{old}} \)) is expected to happen only about 5% of the time by random chance. If we reject the null hypothesis on the basis of these sample means (concluding that the new training system works better), we are running the risk that 5% of the time we will be wrong, because 5% of the time this difference will occur by random chance when the samples are drawn from populations with no mean difference. If \( z = 1.96 \) and we decide the new training system works better (rejecting the null hypothesis), we are running a 2.5% risk of being wrong.
Ok, we have wandered around the question of whether a new HR training program yields higher performance among newly hired realtors than an old realtor training program. We have digested the fact that any actual mean difference observed \((\bar{x}_{\text{new}} - \bar{x}_{\text{old}})\) might be due to the luck of the draw (i.e., random chance) when gathering performance information on 50 folks trained under the new system and 150 folks trained under the old system. We have also discussed statistics that can be calculated \((z\text{ and } t)\) that can tell us how likely any observed sample difference is to have occurred by random chance if there was truly no difference in performance for the populations these samples were drawn from. Finally, we briefly touched on the notion that, while “conventions” exist regarding standards for determining “statistical significance,” these conventions cannot be embraced without consideration of the costs attached to errors associated with such conventions.

Where did \(p < .05\) come from? You might wonder where and how .05, .01, and .001 became the “accepted” levels of Type I error at which folks traditionally concluded two means were “statistically significantly different.” Fischer (1955, 1956) argued that “conventions” such as \(p < .05, .01, \text{ or } .001\) are reasonable when the investigator is starting at ground zero with no prior information about the phenomena under investigation. Unfortunately, social science has evolved in such a way that \(p < .05, .01, \text{ or } .001\) conventions tend to be embraced with little or no thought given to how prior information or relative costs might affect Type I and II error levels. The current debate centers on the fact that cost of Type I and Type II errors are not the same for all null hypotheses, especially in situations where there is at least some prior knowledge about the phenomena under investigation. Costs are certainly not the same when testing the null hypothesis that two sample means are equal in situations comparing two realtor training programs versus comparing the effectiveness of two cancer chemotherapy treatments. Recent controversies alleging the Federal Drug Administration sets statistical standards of “proof” too high before approving experimental drug treatments for humans with advanced prostate cancer ultimately hinge on what are acceptable levels of Type I and Type II error. Costs of Type I and II errors faced in most applied social science work (e.g., human resource management in organizational settings) are assumed to be at levels for which the traditional .05 to .001 standards are acceptable. However, just because some difference is judged to be statistically significant at some “accepted” levels of Type I and II error doesn’t mean it is practically significant. Again, statistical significance reflects a judgment of how likely the observed difference is to have occurred by random chance in samples drawn from populations that did not differ. Examination of Equation 1 should show you that virtually any difference no matter how small can yield \(z = 1.645\) if the samples are large enough! Given a difference is statistically significantly different from 0, “practical” significance reflects a judgment of whether the difference is large enough to matter in some applied context. Conversely, if \(z = 1.6, p \sim .055\) which fails to meet the \(p < .05\) standard and hence is non-significant. Does that mean a huge effect size obtained from \(N = 20\) men in a prostate cancer pilot study should be ignored simply because it barely fails to meet conventional levels of statistical significance? As a current and hopefully long term owner of a prostate, I would vote “no.” More on this (though not on my prostate) latter.
The investigator controls the probability of a Type I error when s/he decides how big a statistic must be for “statistical significance.” Multiplying the probability of a Type I error (e.g., \( p < .05 \)) times its cost gives us an estimate of the expected cost of using \( p < .05, .01, \) versus \( .001 \) as a standard of “statistical significance.” Lower values of \( p \) yield lower expected Type I error costs. Unfortunately, as \( p \) values associated with statistical significance get lower, the probability of a Type II error (failing to reject the null hypothesis when you should have) gets higher. Hence, by adopting a low \( p \) value (e.g., \( p < .001 \)), one decreases the expected cost of Type I errors while simultaneously increasing the expected cost of Type II errors.

Clearly a happy medium would be to set the critical value of \( z \) or \( t \) (i.e., the value of \( z \) or \( t \) that has to be observed for the null hypothesis to be rejected) at a \( p \)-level that minimizes the sum of expected costs of Type I and II errors – note, this is rarely, if ever, \( p < .05, .01, \) or \( .001 \), and is part of the reason why Fischer advocated \( p < .05, .01, \) and \( .001 \) only be used when examining new phenomena on which we have no prior existing information (and hence can’t know the cost of Type I or II errors). We haven’t discussed how to estimate probability of a Type II error yet, though we will later. For now, it is enough to know that an investigator’s prior knowledge of the phenomena under investigation is vitally important when estimating the costs of Type I and II errors as well as the probability of a Type II error. It should be relatively rare that an HR-BOut relationship is being examined in which the investigator has absolutely no insight or expectation regarding what might be observed or the expected costs incurred if one draws an incorrect conclusion from what is observed (e.g., Type I and II errors). Chapter 3 on “Utility” will spend more time on how to estimate both costs and value added by HR systems in dollar terms.

So, we have done it. After calculating the \( z \) statistic found in Equation 1, we either 1) reject our original (null) position that the new training system is only just as good or worse than
the old training system and go forward with continued use of the new training system or 2) fail to reject our original (null) position and realize we have no basis on which to choose between two realtor training systems that are expected to produce comparable business outcomes. Whether we choose the new, old, or both training systems must be decided on some basis other than their expected effects on new realtor performance.

The next section switches gears a little. The section above examined how HR-BOut relationships are detected when one versus another HR intervention is compared. Our BOut measures were continuous, i.e., realtor sales volume, number of sales, and average sale volume could take on any one of a large number of values. In contrast, the HR intervention was relatively discrete in that it could take on only one of two values – either you attended the new or old realtor training program. What if instead of investing in training, our fictitious realty company was considering developing a new realtor selection system? While our BOut measure remains the same (continuous), our HR intervention is no longer discrete. Instead, like the BOut measure, the selection test battery can yield any one of a large number of “scores,” or Xᵢ, for a candidate being considered. The next section talks about how to estimate the strength of a relationship between two continuous measures Xᵢ and Yᵢ.

Correlations and Straight Lines

The word “correlation” is used in many ways in everyday language. Not surprisingly, I will use it in a very specific way here to describe how strong a relationship exists between two continuous measures. Imagine we administered a pilot personnel selection test battery to newly hired realtors on their first day of employment (Xᵢ), then subsequently obtained measures of job performance 12 months later (Yᵢ). “Correlation” will refer to how strong a linear relationship exists between two measures. Whoops – I snuck in a piece of jargon in that last sentence. A
The Power of Straight Lines - Dawes noted the “robust beauty” of simple straight line models a long time ago. For example, when X and Y are related monotonically (i.e., as X goes up, Y always goes up), the relationship could be a straight line relationship or curved relationship where the curve never turns down. Dawes and Corrigan (1974) found that ordinary least squares regression on average is able to predict 92% of the variance in the criterion when X and Y had a curvilinear monotonic relationship. In other words, if one can assume that as X goes up, Y goes up, simple linear prediction models will generally do a very good job of predicting Y relative to predictions that could have been made from the “true” curvilinear model of X→Y relationships if you happened to know what it was (which you usually don’t).

Linear relationship between two variables is characterized by a straight line. Recall the formula for a straight line you probably learned about in 10th grade geometry classes: \( Y = a + bX \), where \( a \) is the Y-intercept and \( b \) is the slope. The Y-intercept is the point where the line crosses the Y axis (where \( Y = 0 \)). The slope is equal to how much change in Y occurs when X is changed by one point (“rise over run” comes from the ratio of how far up or down you have to “rise” with your pencil divided by how far over you then have to “run” horizontally to create a right triangle using the line as the hypotenuse). While you would likely never expect a set of personnel selection test scores (X) and their subsequent BOut measures (Y) to fall exactly on a straight line, it might be of interest to estimate what straight line “best” fits any given set of XY scores. Consider Figure 1 to further set the context – the cigar-shaped ellipse represents a cloud of points, where each point in the cigar represents one newly hired realtor’s overall score on the selection test battery (\( X_i \)) and a subsequent 12 month measure of job performance (\( Y_i \)).

The ellipse or cigar shaped “hot dog” drawn around the cloud of X-Y points visually portrays the general trend or relationship between the two measures. With a little more imagination, one can see the line or “stick” running through “hot dog.” The “best” stick would be one that pretty much goes through the middle of the hot dog, i.e., a stick that minimizes how far each point in the cloud is from the line.
Alternatively, a different “best fitting stick” might minimize the sum of squared differences between each individual point and the line (you knew we would see our old friend “sum of squares “ again!). More on this in a minute.

In Figure 1, “strength” of the relationship is reflected in how lean or fat the ellipse or “hot dog” is around the line or “stick” passing through its middle. Lean, skinny hot dogs mean the relationship is strong, and knowing what someone’s test score $X_i$ is associated with a pretty tight range of possible future job performance levels $s$/he achieves ($Y_i$) - the stick does a reasonably good job of approximating the hot dog. Fat, round shaped hot dogs mean the relationship is weak – sticks don’t do a very good job of approximating meat balls. As we will see below, the narrow definition of “correlation” we will use here directly reflects how fat versus skinny the hot dog is around the “best fitting” stick running through it. Why would you want a high correlation (skinny hot dog)? Because the realty firm faces a pool of applicants with scores ($X$) on the selection test. However, the firm is only truly interested in $BOut$ ($Y$), the job performance measure deemed relevant for its business outcomes. The firm’s only interest in $X$ is to the extent that $X$ helps it predict which applicants will yield the highest $BOut$s. A high correlation (skinny hot dog) between $X$ and $Y$ means that a fairly tight range of $BOut$ $Y$ values occur for each $X_i$. Our best guess of each candidate’s future job performance $Y_i$ will be fairly accurate, because the range of $Y_i$ observed with for all candidates with test score $X_i$ is fairly narrow.

While the correlation, typically written as $r_{xy}$, captures the “strength” of the $X\rightarrow Y$ relationship in our sample, we can also estimate $X\rightarrow Y$ relationship’s “nature,” form, or shape. A formula describing the line (stick) in Figure 1 is generally the best way to do this – for those of you educated in U.S. public school systems, this is the formula you learned for a straight line in
9th or 10th grade geometry, i.e., $Y = a + bX$, where $a$ is the $Y$-intercept and $b$ is the slope. You may recall the fancy statistical version of this formula is $\hat{Y} = \beta_0 + \beta_1 X$, which was called the “regression” line in your prior statistics courses – it is exactly the same formula from geometry class, though some of the symbols are a little fancier.

Where do we get the information needed to draw the hot dog and stick in Figure 1? The best information would come from a sample of, say, $N = 100$ newly hired realtors who:

1. were given the selection test as part of the “normal” application process where realtors are chosen totally at random (silly, I know, but bear with me);
2. were told the selection test would be used in determining whether a job offer would be made (even though it would not be used . . . this ensures each applicant’s test score approximates what it would have been if s/he had taken the test under “real world selection” conditions); and,
3. all survived on the job at least 12 months in order to obtain the same 3 BOOut measures discussed in the realtor training example above.

Now, you are probably wondering what companies are silly enough to hire people based on a random drawing. Most companies would not do this (though unfortunately, I know of some actual examples). For right now, I will only say that by making the 100 candidates hired at random in this example I make sure that any underlying $X \rightarrow Y$ relationship will be the only thing
influencing the correlation. If no X→Y relationship exists, the X→Y correlation should be $r_{xy} = 0$.\textsuperscript{12}

At this point you are probably wondering why we are interested in the correlation ($r_{xy}$) or the formula for the straight line $c$. While this will be discussed in more detail in Chapter 3’s discussion of utility, a quick explanation is probably needed here to keep you all from bailing on me. Specifically, \( \hat{y}_i = b_0 + b_1 x_i \) is the formula that gives the predicted value of Y (represented by a capital letter Y with a “carrot” or hat over it, i.e., $\hat{Y}$) that any applicant with the test score $x_i$ is expected to have. So, if the formula $\hat{y}_i = b_0 + b_1 x_i$ was calculated to predict realtor sales volume, for any future applicant with test score $x_i$, we simply plug her/his score $x_i$ into the formula, which means multiplying it by $b_1$ and adding $b_0$, to yield the estimate ($\hat{y}$) of what that applicant’s sales volume will be in her/his first year of employment. Further, we could plug the average applicant test scores into the formula to estimate what average newly hired realtors’ sales volume will be. We could also do this separately for each applicant recruiting source, which would tell us how much better or worse the average expected dollar sales volume will be for those recruited from expensive versus inexpensive sources of recruits (e.g., on-campus college recruiting versus local newspaper help wanted advertisements). Calculation of the average expected sales volume per recruiting dollar spent on each source would help us focus recruiting efforts on sources expected to yield the highest future sales volume for each recruiting

\textsuperscript{12} In fact, the information that would most likely be available is on a sample of $N < 100$ applicants who met steps 2& 3 (not step 1) after having passed through the firm’s existing selection procedure (e.g., probably some combination of unstructured interviews and maybe someone’s favorite paper and pencil test) and survived their first year on the job. In this instance, the correlation calculated on these $N < 100$ pairs of $X_i$ and $Y_i$ values would likely be smaller (or “attenuated,” the fancy word) than what we would have found if everyone had been randomly selected for employment. If the existing selection system yielded applicant scores $X$ that were at least somewhat related to the new selection system’s $X$ scores, use of the existing selection system would have robbed the new selection test of the opportunity to fully show how strongly it is related to $Y$. If really low performers were more likely not to survive 12 months on the job, turnover would have robbed the new selection test of the opportunity to show it could have predicted these folks would have been low performers.
dollar spent. So, being able to estimate what \( b_0 \) and \( b_1 \) are in the
formula \( \hat{Y}_i = b_0 + b_1X_i \) is necessary if one wants to make important
policy decisions about HR recruiting efforts.

Now, the good news is that virtually no one computes \( r_{xy} \),
\( b_0 \), or \( b_1 \) by hand (i.e., literally or with a calculator) using a
formula, pencil, and paper anymore. Common spread sheet
computer packages (e.g., Excel), advanced statistical software
(e.g., Systat, SPSS, SAS), and some advanced calculators can
give you estimates of \( r_{xy} \), \( b_0 \), or \( b_1 \) from sample data. Formula for
these statistics are presented in virtually all required
undergraduates quantitative methods courses and their associated
text books, so I will not review them here. For purpose of the
realtor selection example, we want to know whether the new
selection test predicts newly hired realtors’ job performance at
the end of their first year of employment. One test of this is
whether the correlation \( r_{xy} \) is statistically significantly different
from 0 (\( H_0: r_{xy} = 0 \) vs. \( H_A: r_{xy} \neq 0 \)). We first put the \( N = 100 \)
randomly hired realtors’ test scores and business outcome data in adjacent columns of an Excel
spread sheet. Selecting any remaining blank cell in the spread sheet, we select formula from the
formula bar that estimate \( r_{xy} \), \( b_1 \), and \( b_0 \). Because \( N > 30 \), the Central Limit Theorem applies,
and we calculate the following \( z \) statistic:

\[
z = \frac{r_{xy} - 0}{\sqrt{\frac{1 - r_{xy}^2}{n-2}}}
\]
Cool Characteristics of $r_{xy}$: $r_{xy}$ has a number of characteristics that make it particularly useful in summarizing the strength of linear $X \rightarrow Y$ relationships. These include . . .

1. $r_{xy}$ has to range in value between -1.0 and +1.0, with larger absolute values describing stronger $X \rightarrow Y$ relationships.
2. If all $X,Y$ pairs fall on a straight line, $r_{xy} = 1.0$ or -1.0 depending on whether $Y$ increases or decreases as $X$ increases.
3. $r_{xy} = 0$ if there is no $X \rightarrow Y$ relationship.
4. The sign of the correlation ($\pm$) will be the same as the sign of the slope of the scatter plot in Figure 1.
5. $r_{xy}$ is “symmetric” in that you get the same value regardless of how the $X$ and $Y$ labels are assigned to the data.

Equation 2

For example, if Excel told us the correlation was $r_{xy} = .30$, then $z = \frac{.30}{\sqrt{1-.09}} = 3.11$. As our observed $z = 3.11$ exceeds the $z$ value from the normal table that cuts off the upper and lower $p = 2.5\%$ of the distribution (i.e., $z = \pm 1.96$), we can reject position that the correlation $\rho_{xy} = 0$ in the population (reject $H_0$: $r_{xy} = 0$, $p < .05$, 2-tailed: note, the “- 0” in the top part of the $z$ formula reflects the fact that we are testing the null hypothesis that $r_{xy} = 0$) and conclude the correlation $r = .30$ is “statistically significantly different from 0 at $p < .05$.” A statistically significant correlation between the realtor selection test and subsequent year one sales volume is evidence of criterion validity, i.e., evidence supporting the inference that performance on the selection test predicts future measures of job performance (one way of establishing job relatedness as outlined in Section 14 of the EEOC Uniform Guidelines on Employee Selection Procedures).

A more likely HR policy scenario would involve asking whether the new realtor selection test incrementally contributes to predicting new realtor performance above and beyond whatever is currently used to select realtors. For example, suppose a structured interview is currently used to score applicants and that the vender we purchased the interview from said showed average correlations of $\sim r = .18$ with subsequent sales volume in other client firm’s newly hired realtors surviving 12 months on the job. One could ask whether the observed correlation between the new selection test and the old structured interview were significantly different, i.e., $H_0$: $r_{new} \leq r_{old}$ vs. $H_A$: $r_{new} > r_{old}$. Unfortunately, things get a little complicated
relative to tests of $H_0: r_{xy} = 0$. In the “Cool Characteristics of $r_{xy}$“ sidebar I noted $r_{xy}$ ranges from -1.0 to +1.0. With a large enough sample size (e.g., $n > 30$) the $z$ statistic in Equation 2 above rapidly approaches normality when in fact $\rho_{xy} = 0$ is true in the population (remember we use Greek letters for population statistics, and $\rho$ is the Greek lower case $r$). Samples drawn from a population with $\rho_{xy} = 0$ have equal “room” to deviate by random chance in either direction, i.e., there is just as much of a possibility that random sampling error could cause a sample’s observed $r_{xy}$ to be greater or less than 0. The numerator of Equation 2 is written $r_{xy} - 0$ to reflect the fact that 0 is the null hypothesized value we are testing $r_{xy}$ against. What if the null hypothesized value of $r_{xy}$ was .18 (the criterion validity estimate we had for the existing structured interview realtor selection procedure)? If the numerator of Equation 2 was rewritten as $r_{xy} - .18$, there is no longer equal room on either side of the null hypothesized $\rho_{xy} = .18$ value for sampling error to occur. By setting $H_0: r_{xy} = .18$, we have effectively imposed a relative ceiling on sampling error that can occur in the positive direction relative to sampling error that can occur in the negative direction. The bad news is that the $z$ statistic found in Equation 2 is no longer normally distributed and cannot be used to test $H_0: r_{\text{new}} \leq r_{\text{old}}$ vs. $H_A: r_{\text{new}} > r_{\text{old}}$. The good news is that Fisher (1915) found a transformation of $r_{xy}$ that he proved would overcome this problem.

Specifically, Fisher’s $z$ transformation:

$$Fisher's \ z_r = \frac{1}{2} \ln \left( \frac{1 + r_{xy}}{1 - r_{xy}} \right)$$

Equation 3

Then, after transforming both the $r_{xy}$ observed with the new realtor selection procedure ($r_{xy} = .30$ becomes $z_{\text{new}} = .3095$) and the old structured interview system ($r_{xy} = .18$ becomes $z_{\text{old}} = .182$), we derive a new test statistic ($z_{\text{test}}$) in Equation 4:
\[ z_{test} = \frac{z_{new} - z_{old}}{\sqrt{1/n - 3}} \]

Equation 4

Equation 4 yields \( z_{test} = \frac{3.095 - 1.82}{\sqrt{1/97}} = 1.26 \), which fails to reject the null hypothesis \( H_0: r_{new} \leq r_{old} \) at \( p < .05 \). Again, this is the test statistic used if we want to know whether an observed \( r_{xy} \) from a single sample of newly hired realtors is significantly larger than some specific non-zero \( r_{xy} \) comparison point. What we were uncomfortable relying on the vendor’s assurances that \( r_{xy} = .18 \), and instead conducted a separate study on another \( n = 100 \) newly hired realtors using the “old” structured interview system? This is called a test of the equality of two independent correlations, simply because each correlation came from two separate sample’s data. The \( z_{test} \) statistic in this instance is found in Equation 5 below:

\[ z_{test} = \frac{z_1 - z_2}{\sqrt{1/(n_1 - 3) + 1/(n_2 - 3)}} \]

Equation 5

There are a number of other circumstances that can cause problems in use if the \( z \) test statistics discussed in this chapter. For example, what if \( z_{new} \) and \( z_{old} \) were derived from the same sample of \( n = 100 \) newly hired realtors? The samples are no longer independent, because the same \( n = 100 \) realtors is used in each sample. Further, the realtor sales volume measure \( Y \) used in deriving \( z_{new} \) and \( z_{old} \) is exactly the same. A number of “work arounds” or adjustments exist to alter or repair various test statistics so questions framed as null hypotheses can still be answered (again, see Bobko, 2001, for examples of tests of equality of two dependent correlations as well as tests appropriate for other circumstances).
A nifty, but brief, detour. Before walking through an example of how correlations and equations for regression lines are used (or not used) to determine whether a selection test is “fair” under the EEOC Uniform Guidelines on Employee Selection Procedures (1978), I want to briefly review one relationship between these statistics that is not always emphasized (and sometimes not even presented) in these courses. Specifically, $b_1 = \frac{r_{xy} SD_y}{SD_x}$, or the slope of the sample regression line is equal to the sample correlation ($r_{xy}$) times the ratio of the sample standard deviation of Y divided the sample standard deviation of X. Similarly, $r_{xy} = \frac{b_1 SD_x}{SD_y}$. What this means is that if you have the sample data and an estimate of $r_{xy}$ (or $b_1$), you can calculate the standard deviations of X and Y and use them in the appropriate formula to estimate $b_1$ (or $r_{xy}$).

This will come in handy latter when we encounter situations where we don’t have either $b_1$ or $r_{xy}$ from a study done within our own firm, but we do have an estimate of $SD_y$ and $r_{xy}$ from studies done with other employers of jobs that are substantially the same as our job. This is useful when you want to forecast future BOut levels (e.g., sales volume, job tenure, etc.) in a job from criterion validity evidence collected from other jobs that are “substantially the same.” The phrase “substantially the same” is the standard specified in Section 7 of the EEOC Guidelines for determining when it is ok to transport criterion validity evidence obtained on one job to another.

An Example of Test Fairness – A Stick Comparison

Regression analysis as described above is the primary tool used to determine whether a personnel selection test is unfair or biased against groups protected by Title VII of the 1964 Civil Rights Act. The following section of the EEOC Uniform Guidelines on Employee Selection Procedures (1978) provides a narrative definition of "test fairness:

29 CFR Ch. XIV (7–1–99 Edition) (a) Unfairness defined. When members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a measure of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores.
As you may be aware, the Guidelines were first published in 1972, with an extensive revision published in 1978. This document is given great deference by the courts in deciding how to rule in any Title VII litigation, and a common pre-emptive defense against such litigation is for organizations to make sure their HR systems are consistent with the Guidelines. One of the most common ways to determine whether a test is "fair" uses what has come to be known as the "Cleary model" or "regression model." Our discussion of correlations and regression above is directly related to this method of determining test fairness, so this seems a good time and place to describe how to determine whether a selection test is “fair” and in compliance with the Guidelines. Consider Figure 2 below in which personnel selection test scores $X$ are plotted against subsequent measures of job performance $Y$ for some groups of recently hired white and black employees. Note the following about Figure 2:

1. Black and white employees on average perform equally well on the job.
2. Black employees perform meaningfully lower on the personnel selection test relative to white employees (by the amount $d = \bar{X}_{white} - \bar{X}_{black}$).

---

13 Perhaps the most easily accessed on-line version of the Guidelines and the subsequent Q & A’s issued by the EEOC to assist in their interpretation are available at [www.uniformguidelines.com](http://www.uniformguidelines.com). While this web site is owned and maintained by a commercial consulting firm (with which I have no ties or affiliation), it is much more “useable” than the materials available at the EEOC official website.

14 One of my former professors, T. Anne Cleary (1968), first described this method almost 40 years ago.
3. For any "cut score" C (i.e., a vertical line drawn from the X axis upwards representing the minimum X score needed to receive a job offer), more white applicants will receive job offers than blacks (both in terms of absolute numbers and as a percentage of those applying).

4. If regression analyses were performed separately for each race, a black applicant who earned score $X_i$ will be expected to generate job performance equal to $\hat{y}_{\text{black}}$, while a white applicant who earned the same $X_i$ score will be expected to generate job performance equal to $\hat{y}_{\text{white}}$.

Note, this can happen because either the slope ($b_1$), Y intercept ($b_0$), or both are different for the black and white regression lines (it happens due to differences in $b_0$ in Figure 2). If the best fitting line capturing the X-Y relationship had been derived on the combined sample of black and white applicants (i.e., the bold line separating the two ellipses), predicted job performance for both applicants would be $\bar{y}$, where $\bar{y}$ actually over predicts true average performance of our $X_i$ white applicant by the amount $\hat{y}_{\text{white}} - \bar{y}$ and under predicts true average performance of our $X_i$ black applicant by the amount $\bar{y} - \hat{y}_{\text{black}}$.

Recall the definition of “fairness” reprinted from the Guidelines above – “When members of one race, sex, or ethnic group characteristically obtain lower scores on a selection procedure than members of another group, and the differences in scores are not reflected in differences in a

How “fair” are things generally? At least two things prevent me from drawing some generalization about “typical” levels of fairness found with different types of personnel selection systems. First, results in any given situation are influenced by many factors, with things like aggressiveness of minority recruiting, depth and quality of the relevant applicant pool, and type of performance measure used (e.g., absenteeism, job tenure, supervisory performance ratings) affecting whether fairness is present. Second, few if any of these types of studies ever get published in the public domain, especially if they show an absence of fairness. Most are found in proprietary technical reports deep in the bowels of HR departments and consulting firms.
measure of job performance, use of the selection procedure may unfairly deny opportunities to members of the group that obtains the lower scores.” Figure 2 shows a “classic” case (as originally described by Cleary, 1968) of unfairness, as the black applicant group scored meaningfully lower than the white applicant group. Some of you may note that in addition to being unfair, use of the cut score C in Figure 3 will likely cause “adverse impact,” i.e., the proportion of blacks hired relative to the number of black applicants will be less than $4/5$ of the proportion of whites hired relative to the number of white applicants, or $80\% > \frac{N_{\text{blacks hired}}}{N_{\text{blacks applied}}} \div \frac{N_{\text{whites hired}}}{N_{\text{whites applied}}}$. This shows that adverse impact can occur in the presence or absence of the test being fair.\(^{15}\)

Hopefully is should be clear that a perfectly “fair” personnel selection test that has no adverse impact would be one where the black and white “hot dogs” perfectly overlap on the same “stick.” In this situation, every applicant regardless of race who earns a test score $X_i$ is expected to yield the same level of $B_{\text{Out}}$, $\hat{Y}_i$, and no adverse impact is expected to occur regardless of where the cut score C might be drawn. Unfortunately, “fair” tests do not always behave this way.

Figure 3 shows an example of a “fair” test under the Guidelines that is also likely to have adverse impact. In this instance, differences in typical scores on the selection tests are reflected in differences in job performance. Again, the test is fair because the hot dogs are on the same stick, i.e., the regression coefficients estimating the slope ($b_1$) and $y$-intercept ($b_0$) of the best fitting line through the combinations of test scores and subsequent job performance ($X,Y$ points

\(^{15}\) Adverse impact is not illegal. However, if it is present the Guidelines describe a whole bunch of evidence the organization is expected to provide to justify use of the test. Evidence showing “test fairness” is one of those pieces of information. Job relatedness, as evidenced by a significant criterion validity correlation coefficient, is another.
in the ellipse) are the same for the white and black applicant pools. If you recall that Title VII of the 1964 Civil Rights Act applies to race, sex, creed, color, and religion, the example in Figure 3 might be compelling if you replaced “white” with “female” and “black” with “male,” while making $X_i = \text{scores on a finger dexterity test}$ and $Y_i = \text{small parts assembly task performance on a computer component assembly line}$. Research in fact shows men on average demonstrate lower levels of finger dexterity ($X$) than women and tend to correspondingly perform lower on jobs tasks requiring finger dexterity ($Y$) (Peters, Servos, & Day, 1990; Thomas & French, 1985).

Note for a moment that differences in correlations ($r_{xy}$) between groups (black/white, male/female, etc.) does not constitute evidence of an unfair test. When this occurs it is typically called “differential validity,” and a lot of time was spent back in the 1970’s sorting out why comparisons of $r_{xy}$ across protected subgroups is not a good way of measuring “fairness.” Those interesting in reliving these exciting times are invited to read Bobko and Bartlett (1978) and Bartlett, Bobko, Mosier, and Hannan (1978) for a thorough discussion of the issues.

Ok, now that we know what a “fair” and “unfair” personnel selection test looks like, how do we know whether we have one or not? By drawing an inference from a statistic, of course! Since the formula $\hat{Y}_i = b_0 + b_1 X_i$ describes the defining $Y$-intercept and slope characteristics of the

**Figure 4: A fair or unbiased test that has adverse impact (or 2 hot dogs on the same stick)**

- $\hat{Y}_i = b_0 + b_1 X_i$
- $b_0, b_1$ are the intercept and slope, respectively.
- $X_i$ and $Y_i$ are the scores on the test and job task performance, respectively.
line, one might suspect some statistic is sensitive to whether $b_0$ and $b_1$ are the same for minority group members and majority group members. What would be best, if possible, is a statistic that simultaneously tested whether $b_0$ and $b_1$ are significantly different across relevant majority/minority groups.

In fact, there is! T. Anne Cleary (1968) described how. Specifically, consider the following four prediction equations:

\[
\hat{Y} = b_0 + b_1 X_{test} : r_{Y,X_{test}}^2 \\
\hat{Y} = b_0 + b_1 X_{race} : r_{Y,X_{race}}^2 \\
\hat{Y} = b_0 + b_1 X_{test} + b_2 X_{race} : R_{Y,X_{test},X_{race}}^2 \\
\hat{Y} = b_0 + b_1 X_{test} + b_2 X_{race} + b_3 X_{test} X_{race} : R_{Y,X_{test},X_{race},X_{test}X_{race}}^2
\]

Equations 6-9

We will not explore all the nooks and crannies of these four equations right away, as there are many and they will reveal themselves over the remainder of the book. Right now I ask that you pay attention to only four things about Equations 6-9. First, Equations 8 & 9 represent multiple regression models, while Equations 6 & 7 are called simple regression equations (simple regression involves one predictor, multiple regression uses more than one predictor). Simple regression models are easy to draw in 2-dimensional space (e.g., see Figure 1 above). Multiple regression models require $k + 1$ dimensional space, where $k$ is the number of predictors (the +1 is for $Y$). Sorry . . . while technologies exist for artfully drawing 3-dimensional (especially when one variable only takes on two possible values, as in Figure 3), I am unaware of technologies that yield visualizations of 4- or more-dimensional models. Second, $X_{race}$ is coded so Caucasian = 1 and African American = 2 (or vice versa). Dichotomous variables (i.e., variables that take on only one of two values) like sex or race (when comparing a minority group to a majority group) are commonly called a “dummy” variables, a label used when a measure
contains a small number of discrete values (a continuous measure contains a large number of possible values). Third, the equations all end with a colon followed by a “coefficient of determination,” either an r² (signifying simple regression with one predictor) or “multiple” R² (with the capital R signifying multiple predictors). The r² and R² subscripts can be read as “r² or R² when Y is regressed onto . . . {list of predictors used in that equation}.” Coefficients of determination are simply squared correlation coefficients and represent the proportion of variance in Y explained by the model’s best prediction (i.e., \( \hat{Y} \)). Finally, fourth you should notice Equation 9 contains three predictors, with the last one formed by multiplying the first two, i.e., \( X_{\text{test}}X_{\text{race}} \).

Reverting back to our realtor selection example, let’s assume the sample \( n = 100 \) consisted of \( N = 30 \) African Americans and \( N = 70 \) Caucasians. Cleary (1968) showed how z statistic testing the null hypothesis \( H_0: b_3 = 0 \) in Equation 9 simultaneously is sensitive to whether \( b_0 \) and \( b_1 \) are the same when Equation 6 is derived separately for the \( N = 30 \) African American realtors versus the \( N = 70 \) Caucasian realtors. Recall the z statistic testing \( H_0: r_{xy} = 0 \) in Equation 6 will tell us whether the selection test has predictive criterion validity (\( r_{xy} \)). Similarly, the z statistic testing \( H_0: r_{xy} = 0 \) in Equation 7 will tell us whether \( \bar{Y} \) differ for African American and Caucasians in this sample (this is due to the “dummy” coded nature of \( X_{\text{race}} \)).

Again referring to an Excel spread sheet containing columns with \( X_{\text{test}}, X_{\text{race}}, X_{\text{test}}X_{\text{race}}, \) and \( Y \), I would find open cells in which I would use the regression estimation tools in the formula box to estimate \( b_0, b_1, b_2, b_3, \) and \( R^2 \) for Equation 9 to determine whether the test was racially “fair” in this sample. Excel is nice enough to generate estimates of \( b_0, b_1, b_2, b_3, \) and \( R^2 \) for Equation 6 and estimates of their standard error (the denominator of the z statistic testing the five hypotheses \( H_0: b_0, b_1, b_2, b_3, \) or \( R^2 = 0 \)) and the z statistic and its probability of occurrence (p value). Hence,
one simply looks across the screen to see whether the p value associated with the z statistic reported for $b_3$ is less than .05. If it is, $H_0: b_3 = 0$ is rejected and we conclude the selection test does not meet the EEOC Uniform Guideline for Selection Procedures for fairness. If $H_0: b_3 = 0$ is not rejected, we conclude the selection is fair.

A couple of real world “gotchas” need to be mentioned here. First, recall that Title VII of the 1964 Civil Rights Act establishes protected employment rights for different races, creeds, colors, national origins, sex, and religions. This law exists because some of these categories constitute distinct minorities within the U.S. population that have, historically, received different treatment in the market place and society as a whole. By definition, “minorities” occur with less frequency than “majority” group members in the population. Hence, sample sizes used to determine things like test fairness can be a problem. Notice in the realtor selection example above I assumed $n = 30$ African American and $n = 70$ Caucasians made up the original $n = 100$ realtors tested. I made this assumption so the Central Limit Theorem would apply regardless of how the data might be racially grouped. However, actual samples of realtors going through such a system might only contain 5 African Americans, making it impossible to tell whether the selection system has adverse impact or whether it exhibits fairness (the courts ignore the $4/5$ths rule and generally don’t attend to adverse impact when the change of one or two individuals in the numerator or denominator of the utilization ratio radically alters whether the $4/5$ths rule is met). The last time I looked, 58% of all realtors were women, so I would expect little if any difficulty when examining whether a selection test is gender fair.
A second “gotcha” occurs when one finds evidence that a selection test is not fair even though it has high criterion validity ($r_{xy}$ in Equation 6). The firm is faced with a selection test that does a real good job of predicting performance even though it does so in an unfair way. Ouch . . . is there any way to “fix” the test that will make it fair and still be able to predict applicants’ future performance? The short answer is “yes,” though it is illegal! Recall in Figure 2 above I said $d = \bar{X}_{\text{white}} - \bar{X}_{\text{black}}$. If we added this difference score “d” to all black applicants’ test scores, the black ellipse in Figure 2 would move to the right until it overlapped with the white ellipse, and both hot dogs would be on the same stick (this has the same effect as subtracting “d” from each white applicants selection test score, moving the white ellipse left to overlay the black ellipse). Alternatively, we could administer the test to all applicants, then do “top down selection” proportionately within race, offering jobs to the highest scoring black and white applicants in a way that 3 blacks receive offers for every 7 whites receiving offers.

Collectively, these approaches are often labeled “race norming.” Curiously, for a while in the late 1980’s the federal Office of Personnel Management race normed scoring of the General Aptitude Test Battery (GATB), the test administered by all public employment agencies nationwide. I have never seen an

**More on Race Norming:** An example makes it clear why the legislature (in the 1991 Civil Rights Act) and courts have a negative view of race norming. Imagine a black and white applicant sitting in a public unemployment agency taking the General Aptitude Test Battery (GATB). The candidates possess equal knowledge, skills, and abilities as tested on the GATB. In fact, they are so similar in our fictitious example that they give identical answers to each GATB question. After sitting for the test, the applicants relax over coffee and share their impressions of the test, where they are both surprised to learn they answered the questions in exactly the same way. Sometime latter both candidates return to learn how they scored. Imagine their surprise when they learn the white applicant’s score was X, while the black applicant’s score was X+d! This oddity occurs because race norming uses information gleaned about group differences to make adjustments in individuals’ test scores. Clearly most black and white applicants must choose different answers to GATB questions, otherwise the difference in group means $\bar{X}_{\text{white}} - \bar{X}_{\text{black}} = d$ would be 0. The example above illustrates a possibility that could nonetheless occur and what most people can’t help but consider fundamentally unfair, i.e., a situation where two individuals “earned” the exact same score on a test, though the white applicant’s score is “race taxed” by the amount $–d$. Unfortunately, situations like this occur with alarming regularity whenever information gathered at one level of analysis (e.g., racial group) are used to draw policy implications at another level of analysis (e.g., the individual). This exact situation prevents race norming from being an “easy fix” for otherwise criterion valid personnel selection tests.
explanation of why this occurred, probably because the first President George Bush temporarily suspended use of the GATB once he learned of it (once the race norming step was removed, test administration resumed). Race norming, while not explicitly illegal under Title VII or deemed inappropriate by the Guidelines, is clearly an example of disparate treatment, since one protected group is getting “d” points added to their individual scores just because they are members of that protected group. Regardless, the 1991 Civil Rights Act amended Title VII of the 1964 Civil Rights Act to explicitly make race norming illegal. So, while it is very tempting and appropriate from a test developer’s point of view to find ways to modify the selection test so as to retain criterion validity and eliminate unfairness, race norming is not permissible under the law.

Finally, Figure 4 shows a test that meets the EEO Guidelines fairness criterion yet exhibits differences in the strength of the X-Y relationship for black and white applicants. This was commonly referred to as "differential validity" during the 1970's, when conceptual definitions of test fairness were being thrashed out in the literature. It was not deemed as good a metric of test fairness as the Cleary model and has not been embraced in the literature or courts. See original research by Bartlett, Bobko, Moser, and Hannan (1978), Bobko and Bartlett (1978), and Arvey and Faley (1992) for a sampling of studies addressing these issues.

At the risk of running the “hot dog on a stick” metaphor totally into the ground, Figure 4 is perhaps best conceived in terms of an Oscar Meyer product called a "cheese dog," i.e., a hot
dog product (the outer ellipse) which has been injected with an inner "ellipse" of cheese. If you haven’t realized it by now, I am not picky – any metaphor to help the reader create a “hook” to hang the information from is alright by me.

**Conclusion**

Well, hopefully that wasn’t too painful a trip down your statistical memory lanes. We started with a few real HR policy/practice questions and worked backward to try and get an intuitive understanding about why the following three statistical tests are relevant:

1. tests of whether two group means are significantly different;
2. tests of whether a correlation is significantly different from 0; and,
3. tests of whether regression models ($b_0$ and $b_1$, or Y-intercepts and slopes) of selection test score relationships to BOut are the same for two groups protected under Title VII.

Hopefully the readers’ comfort levels have (at least slightly) incrementally increased in the use of these statistical tools relative to comfort levels typically achieved in required statistics courses.

To be sure, there are a lot of additional statistical tools that prove useful to have in your box. Fortunately, there is also usually someone you can find who has a deeper/broader tool box and knowledge of the tools in it than you. It might be helpful to view this aspect of your HR responsibilities like you view MIS – a necessary evil. I personally don’t learn anything about a computer unless I have a need for it today, as if I need it tomorrow it will surely change between then and now, causing me to waste time learning unnecessary stuff. Once you attain a minimal comfort level with some basic kinds of statistical inferences, it is (in my opinion) most important to simply stay motivated and open to the fact that you will certainly have to tap other statistical resources (e.g., software, consultants, junior colleagues with “fresher” quantitative training, etc.) to address the various convoluted HR problems and questions that might present themselves. I
have been intimately involved in applying statistical tools to address HR problems for more than 30 years, and I only know of 3-4 individuals who I consider versed in almost every statistical tool out there. To beat up a different metaphor for a while, these individuals are “tool and die” engineers who have intimate knowledge of the inner mathematical workings behind these statistical tools and appreciate the substantive management issues to which they are applied.

Some of the rest of us might be considered “master mechanics.” This book was written because too many HR professionals cannot change their own oil. While they are perfectly capable of routine maintenance and monitoring activities (e.g., changing oil, spark plugs, etc.), they just don’t want to because of a bad first experience under the car.

Ok, I’ll stop. The next chapter is the last “tool” focused chapter, dedicated to introducing a way of estimating the impact of HR policies and practices on the firm using BOut metrics that line managers are responsible for (e.g., dollars). Yes, there is a way of translating criterion validity into dollars and it has been around since the late 1940’s. Unfortunately, a combination of circumstances (e.g., HR professionals’ dislike of statistics and inability to estimate one part of the model until the early 1980’s) has prevented its wide spread use in industry. Hopefully Chapter 3 will make a dent in changing that.
References


Peters, M., Servos, P, & Day, R. 1990. Marked sex differences on a fine motor skill task disappear when finger size is used as covariate. Journal of Applied Psychology, 75, 87-
90.
