MA 2611 Applied Statistics

Christopher Myers

June 13, 2020

Contents

1 Intro

Statistics is the study of data, or the study of pieces of facts or information, such as yearly average global surface air temperatures, the lifetime of electrical transformers, the breaking strengths of metal pins, or the percentage of lymphoma patients treated at medical centers that survive more than 5 years after diagnosis. This sort of information is useful when making determinations about a population, or when making decisions in which data is useful.

In statistics, a sample of a population is used to draw conclusions about the entire population. In order to gather valid data in a scientific way, the data must be selected in a statistically valid way. Random selection isn't always best, as the results won't be completely random (there is a certain amount of quantifiable regularity).

There are four main objectives to this course:

1. **Produce data:** obtain data in a valid way
2. **Summarize data:** appropriate graphical and numerical summaries of data
3. **Estimation:** use data to estimate population quantities of interest
4. **Hypothesis Tests:** use data to test hypotheses

2 Producing Data

Fancy analysis cannot fix a poorly designed study, so collecting data correctly is critical. A study is an examination of a subject for the purpose of advancing knowledge. Many studies require analysis of data.

Available data is data that was obtained prior to the study, but not with the study specifically in mind. There is no guarantee that results based on such data are scientifically valid. Drawing definitive conclusions from available data is often questionable, but using available data for exploratory analysis is an important part of the scientific/research process nonetheless.

Statistically designed studies obtain data used a plan, ensuring that specific points of interest are answered in a scientifically valid way.

Example: Bronze bushing roughness data A manufacturing company found that surface roughness varied too much. To identify some possible causes, the company conducted an exploratory analysis of historical data, and found four options: lathe operator, cutting speed, feed rate, and tool type.

Using this information, the company designed a study that purposely varied these elements to determine what the most influential cause was.

2.1 Types of studies

The main types of studies are controlled experiments, observational studies, and sampling studies. Are use observational units, or entities on which measurements or observations can be made. These observational units are often selected from a larger population.

2.1.1 Controlled Experiments

In controlled experiments, an important element is a *factor*, or something thought to influence the response. Experimental factors are purposely varied by the experimenter, while other unvaried factors are called nuisance factors. Levels are a value assumed by a factor in an experiment. Treatment are Combinations of levels of experimental factors for which the response will be observed.

Controlled experiments are then studies in which **treatments are imposed on experimental units** in order to observe a response. Effects are the differences between the means of different levels that the study tests, and are ultimately what the study is after.

Factors in an experiment are said to be **confounded** if their individual effects cannot be separated. This isn't a good thing.

The two most compelling reasons to conduct a controlled experiment are to provide the foundation for statistical inference and to establish causality. To show causality means to establish that a change in a given treatment will change a given response. Using randomized controlled experiments can help with this by "washing out" the effects of unanticipated nuisance factors. When treats are assigned to experimental units at random, the study is called a randomized controlled experiment.

There are two major methods for assigning treatments: completely randomized design (CRD) and Randomized Complete Block Design (RCBD). Completely randomized design does what it says on the box and works best when experimental units are generally homogeneous, while RCBD groups experimental units into blocks, then randomly assigns treatments within those blocks. This makes RCBD more effective where units within the block are more homogeneous than all units put together. These two methods have similarities to simple random sampling and to stratified random sampling, but their purposes are very different. The sampling methods are to designed to obtain samples that represent a population, while treatment assignment methods are used to remove bias and help demonstrate causality.

There are a few elements of good practice to consider here: For one, block what you can, randomize what you cannot. This helps eliminate nuisance factors. Replicate your data collections as time and budget permits. Replication means repetition, but beware of duplication. Finally, make sure to confirm the results — for example, run some confirmatory experiments.

Experiments on Human Subjects When using humans as subjects, there are some new terms:

- **Treatment group** — group of subjects that receive a treatment
- **Control group** — a group of subjects that receives either no treatment or a neutral treatment.
- **Placebo** — a neutral "fake" treatment given to subjects in a control group in order to counteract the placebo effect (effect that causes people to exhibit the effects of a treatment simply because they believe they received a treatment that would cause it).
- **Double blind** — an experiment where neither the test subjects nor the researchers administering the experiment which group each subject is in. This is used to eliminate both bias and the placebo effects.
- **Pairing/Matching** — a form of blocking, subjects are matched together on the basis of nuisance variables like age, gender, health, etc.

To date the largest experiment ever run on human subjects is the Salk Vaccine Field Trial, the trial that proved the effectiveness of the polio vaccine.

2.1.2 Observational Studies

Observational studies are studies where no actual conditions are modified by researchers; there is no concept of an independent or dependent variable in these studies. Instead, researchers go out into the field and gather data from current events or specific situations.

Cohort/Prospective Studies Also known as quasi-experiments. They don't control the assignment of experimental units, so they can only demonstrate association (correlation) and *not* causation. For obvious reasons these types of studies are not ideal, but they are useful in scenarios where controlled experiments are not possible, e.g. due to ethical or practical concerns.

There are still groups, however, the "treatment" and "control" groups, known as cohorts. These are based on the hypothesized cause, and pattern of response (effect) is then compared between the groups. For example, in a study about smoking causing cancer, you could choose cohorts of smokers and non smokers, then check the incidence of lung cancer between the two groups. If one group gets more cancer than the other, you can conclude there is an association between smoking and cancer.

Case-Referent/Retrospective Studies In these studies, groups are formed based on the response and patterns in hypothesized causes are compared from group to group. In another smoking/cancer example, you might form two groups, one containing people with lung cancer, and the other containing only people without lung cancer. Then conduct an analysis to see which group has more smokers in it, and determine an association based on that.

Note that in retrospective studies you **must** know the effect first. Regardless, they are still useful when the time between hypothesized cause and observed effect is large, or when the effect is very rare. For instance, studies of a rare disease might use retrospective studies.

It is a misconception that studies using past collected data or historical data are retrospective studies by nature. Retrospective/case-referent only refers to the fact that groups are formed based on effect, and not based on cause.

2.1.3 Sampling Studies

Sampling studies rely on data obtained by sampling from a larger target population. Their goal isn't necessarily to establish association or cause/effect, but to infer some status of the population based on what is observed in the sample.

2.2 Selection of sampling units

To properly select sampling units, you need a sample that mimics its origin population, although on a smaller scale. You also need to quantify how far from the population the results are likely to be.

Terminology:

- Target population: a collection of sample units that data is to be gathered from
- Sampling design: a pattern, arrangement, or method used for selecting a sample of sampling units from the target population.
- Sampling plan: the operational plan, including the sampling design, for actually obtaining or accessing the sampling units for the study.

Example: Election sampling In order to estimate the "population" of WPI students prior to an election, select 20 WPI students and interview them. For this study, the sampling unit is an individual student, the target population is all WPI students, the frame is the campus directory, the sample is the 20 selected students, the sampling design isn't chosen yet (too many choices), and the sampling plan is a plan for the implementation of taking the study.

2.3 Sampling Methods

Simple Random Sampling Select samples randomly. Each possible sample then has the same chance of selection. This method is good if sample units are homogeneous and easily accessed.

Stratified Random Sampling Divide the population into different distinct groups (strata), then perform a simple random sample on each of those group. This is more complicated than simple random sampling, but it yields more accurate results in certain circumstances, notably in cases where units in each stratum are more homogeneous than the whole population, or if you want to include small subgroups in the sample.

Cluster Sampling Units in close proximity are grouped in clusters, which are then sampled instead of the individual units. This is useful in cases where going to random locations to obtain data may be more costly — for instance, sample clusters of a population (ex. neighborhoods).

Many samples are taken in stages, using any of the above methods in any stage. For instance, a monthly study conducted by the US Census Bureau estimates values such as unemployment and schooling by taking a multistage cluster sample of 100,000 people in 60,000 households.

2.4 Error

There are two types of errors: sampling error and non-sampling error. Sampling error is just the result of the sample not being a perfect representation of the population, and is to be expected. Non-sampling error can be avoided, however, and comes from various sources, including:

- Inability to sample entire population
- Inability to sample from some units
- Invalid, misleading, or false measurements from some selected units.
- Selection bias — introduced to sample because the sampling method misses certain segments of the population.

Additionally, if sampling units are selected by some non-probability method (e.g. convenience sampling), the results of the study may not be generalizable to the population.

3 Basic Data Analysis

The general idea to data analysis is that data has variation, but there is still an underlying pattern (at least in cases where there is some correlation). By analyzing that pattern, you can figure something out about process or population that the data came from.

3.1 Plots

Frequency Histogram Displays a static pattern of variation without graphing the data according to time. For instance, a graph of temperature anomalies won't show temperature anomalies over time, it will show how often each anomaly value occurred (usually by dividing them into discrete "buckets" of values, e.g. 0.0 to 0.2). Called a frequency histogram because the vertical axis is frequency, not some other unit.

When analyzing frequency histograms, there are a number of things to look for:

- Modality — how many peaks does the graph have?
- Symmetry — is the graph seemingly mirrored about some vertical line?
- Center — does the graph have a center? If so, where is it?
- Spread — how widely spread out is the data?
- Patterns and deviation — what are the main patterns? How does the data deviate from them, if there are any patterns and there are actually deviations?

Note that since frequency histograms don't show time, they cannot be used to infer a relationship between the passing of time and the variance of some value.

Time Series Plot Also known as a line plot. Shows a pattern of variation over time, with the measured values in the vertical column and time in the horizontal column. These graphs can be used to show a relationship between time and a variable.

A process is stationary if the pattern of variation does not change as more data is collected. To assess stationarity, data must be plotted against time. If data is stationary, plots displaying static patterns of variation over time (like frequency histograms) are okay, but otherwise it's better to use another type of plot.

Stratified Plot Stratified plots show data when broken down into subgroups (strata). These are best used to show differences between the strata. For instance, if you were to conduct a study to determine the average weights of bread produced by three different bread machines, a stratified plot would be best because it would show you the performance of each machine individually.

These plots usually have the strata in the horizontal column and the data scale in the vertical column. Individual data points are graphed as actual points above their respective strata labels in a very narrow column.

Everything from here on out is about stationary data. Before applying any of these plots/analysis methods, check for stationarity!

Bar Chart No real explanation needed here. Bar charts are good for showing the size of individual categories or strata. There is one bar for each categorical value, while the vertical axis shows the size of that category.

Needle Plot Used to show a small number of data points (usually 20 observations or fewer) of quantitative data. These are a bit like bar charts, but they operate over a small set of discretized categories. For example instead of a pass/reject/rework category system for measuring diameters on a bar chart, you could make a needle chart showing those diameters on the horizontal axis.

Note that too many "categories" drastically lowers the usefulness of the chart. In such cases it's better to use a frequency histogram instead.

3.2 Sum of Squares

Sum of squares is a way of looking at the variation of a data series. Basically, calculate the average of the data set (or stratum, if applicable), subtract that from each data point, and square the results. Then just add all of those together, and you have a sum of squares. If there are multiple strata, you can calculate the within variation (variation within a stratum) and variation between (variation between strata). In the case of within variation, use the average of stratum, but for between variation, use the average of each stratum's data (?).

3.3 Variation

There are two major sources of variation in a measurement (or gage-ing process): repeatability and reproducibility. Repeatability is the consistency of the gage (measuring tool), while reproducibility is whether the same measurement can be obtained using the same tool or not (?).

For example, in a gage R&R study, four operators of a laser ranging device each took fifteen measurements of the same distance. The results were stratified by operator, and showed no pattern between which measurement in sequence was taken. In this case, the repeatability is the amount of consistency across a single operator's measurement, while the reproducibility is the consistency of measurements taken by different operators. These correspond to within variation and between variation, respectively.

There are some useful numbers involved in variation and basic data analysis. Assume data points $y_1, y_2, \ldots, y_n$:

- Mean: $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$.
- Median: Q_2, the data value in the middle of the set. $y_{n/2}$.
- Mode: The location of the modal bar on a frequency histogram, i.e. location of highest frequency.
- Quantiles: For a number q between 0 and 1, the q^{th} quantile is a value at or below which a proportion at least q of the data lies at or above which a proportion at least $1 - q$ of the data lies.
- Quartiles: A set of particular quantiles, set at $Q_1 = .25, Q_2 = .5, Q_3 = .75$. Note that Q_2 is the median.

Note that for mode, there are different ways of looking at it. It's often more useful to calculate mode based on the bars in a frequency histogram than it is to look at individual data points, as the former will usually yield more useful information. Anyways, more measures:

- Mean absolute deviation: The "average' distance from the mean, $\frac{1}{n-1}\sum_{i=1}^{n}|y_i - \bar{y}|$. There's an $n-1$ out front in the formula because there's really only $n-1$ pieces of information; that... actually, nevermind, I don't get it. Something here about degrees of freedom (see: future) and that $\sum_{i=1}^{n}(y_i - \bar{y}) = 0$.
- Standard deviation: Square root of the average squared distance from mean: $\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2}$. This is also known as the root mean square, or RMS.
- Interquartile Range: The spread of the middle 50% of the data: $Q_3 - Q_1$.
- Outliers: extremely unrepresentative data values. These can be "bad" data values, or they might be good but unexpected values. In either case, outliers should be identified and checked. Box & whisker plots are good for this.

In many cases, a data set can be summarized by five numbers (the "five number summary"), composed of the three quartiles and the upper/lower adjacent values A_+ and A_-. A_- is the smallest data value greater than $Q_1 - (1.5)(IQR)$, and A_+ is the largest data value smaller than $Q_3 + (1.5)(IQR)$.

Example Calculate a five number summary for the data points $240, 144, 167, 172, 143, 133$.

First calculate the quartiles. Sort the data from smallest to largest. If there are an odd number of values, the median Q_2 is the middle data point. Otherwise, the median is halfway in between the two middle values. Q_1 is then the median of all data values less than or less than/equal to Q_2, whichever is easier compute. Q_3 is essentially the same, except for all data values greater than Q_2.

Applying this, the ordered data is $133, 143, 144, 167, 172, 240$. There are six data points, so $Q_2 = \frac{144+167}{2} = 155.5$. The values less than Q_2 are 133, 143, and 144, so $Q_1 = 143$. Doing the same thing on the other side yields $Q_3 = 172$. The IQR is $Q_3 - Q_2 = 29$, so the lower adjacent value 133 and the upper adjacent value is 172 (even though 172 is Q_3).

Summary measures like these can be considered resistant if they are not seriously affected by outliers. Median and IQR, for instance, are resistant, while mean and standard deviation are not resistant. Although mean isn't resistant, it often is a better measure of location than median, as it used the data more efficiently.

There are ways to improve the mean's resistance, however, by using either the trimmed mean or the Winsorized mean. The k-times trimmed mean omits the k largest and k smallest values and computes the mean using the rest. The k-times Winsorized mean performs a similar operation, by creating a new data set by replacing the k smallest data values with the value of the $k+1$st smallest, and the same thing in reverse for the k largest values. The remaining data values are left untouched in the new data set, which is then used to compute the final mean.

4 Statistical Inference

Statistical inference is the process of using data about a sample to draw conclusions about the population. It's easy to take a sample, run an experiment (or take some observations), and run an analysis of that data, but how well do the conclusions apply to the rest of the population? There will undoubtedly be some error involved, so how big is that error? These are the kinds of questions that statistical inference is involved in.

Suppose a study were done on a medication that's supposed to reduce cholesterol levels. A sample of 10 patients is used, and their levels are measured before and after the treatment. Data points caclulated from the sample data are referred to as *estimators*. The sample mean is denoted by $\bar{y}$. The sampe proportion could be denoted by $\hat{p}$; both of these single values are called *point estimators*.

From a set of data (not shown here, it's an in-class example that I didn't copy down), we could conclude that the sample proportion (the number of patients who were correctly affected by the treatment) is $\hat{p} = 0.9$, and the sample mean is $\bar{y} = 21.19$. This doesn't tell us how close those values are to the true population values, however; for that we need the sampling distribution.

The sampling distribution of an estimator arises from the idea that we obtain the subjects in the study by sampling randomly from the population. If you take multiple samples, you can calculate $\bar{y}$ and $\hat{p}$ for each different sample. The sampling distribution is the pattern of variation shown by the values obtained when the estimator is calculated for all possible samples. In the cholesterol example, the sampling distribution of $\bar{y}$ would be the set of values of $\bar{y}$ computed from all possible samples of size 10. If you were to graph the distribution, it might look something like a bell curve, for instance.

If your sample size is 10, it can be shown that the sampling distribution of $\bar{y}$ has mean μ and standard deviation $\sigma/\sqrt{10}$. That is to say, the sampling distribution is much more spread out than the distribution of the population. The standard deviation of the sampling distribution of an estimator is called the standard error of the estimator, so the standard error of $\bar{y}$ is that $\sigma/\sqrt{10}$ value.

To standardize all the $\bar{y}$ values in the sampling distribution, subtract their mean μ and divide the result by their standard error. The resulting values will have a distribution that has a mean of 0 and a standard deviation of 1.

If you know the sampling distribution, you can mathematically determine the percent of samples that will contain μ (the actual population mean). For example, for 95% of samples using the previous data:

$$\begin{aligned} 0.95 &= Pr(-1.96 < \frac{\bar{y} - \mu}{\sigma/\sqrt{10}} < 1.96) \\ &= Pr(\bar{y} - 1.96\frac{\sigma}{\sqrt{10}} < \mu < \bar{y} + 1.96\frac{\sigma}{\sqrt{10}} \\ a &= \bar{y} - 1.96\frac{\sigma}{\sqrt{10}} \\ b &= \bar{y} + 1.96\frac{\sigma}{\sqrt{10}} \end{aligned}$$

This process will give you the values for a and b. No, I don't know where the 1.96 came from, not yet anyways. This interval would be called a 95% confidence interval. Because they give a range of likely values for what is being estimated (here, μ), confidence intervals are examples of *interval estimators*. This particular interval estimator is much more informative than the point estimator $\bar{y}$ because you can figure out $\bar{y}$ if you know the interval (it's at the center!), and the interval gives a range of likely values for μ based on the variation in the sampling distribution of $\bar{y}$.

Back to the original data stuff. The mean of the values is 21.19, and assume that we know the population standard deviation σ is 16 (this is a mystery). A 95% confidence interval for μ is (11.27, 31.11). This is actual statistical inference at work: we've determined (semi-magically) that an acceptable range for μ is from 11.27 to 31.11.

4.1 Normal Distribution

The Normal (or Gaussian) distribution is the most frequently used model for continuous data (such as LDL decrease from the above example). The normal model is given as:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp(-\frac{1}{2}[\frac{x - \mu}{\sigma}]^2)$$

This equation is valid for $-\infty < x < \infty$, and produces a well-known bell curve shape. The standard normal curve has a mean of 0 and a variance of 1, so you can build any arbitrary bell curve using just those values μ and σ. μ controls the position of the curve along the X axis, while σ changes the shape (a smaller σ results in a narrower curve).

This curve has a number of properties:

- The curve is unimodal and symmetric about μ.

- For any $a < b$, the area under the curve between a and b is the proportion of the population values falling between a and b.
- The total area under the curve is 1.
- If a population of values follows an $N(\mu, \sigma^2)$ distribution, and if we standardize each value by subtracting the mean μ and dividing the result by the standard deviation σ, the population of values follows the standard normal distribution of $N(0, 1)$.

This is where the magical values of -1.96 and 1.96 came from earlier. The area under the curve from -1.96 to 1.96 is 0.95, assuming that the LDL data was normally distributed.

Example Specification limits for the diameter of a nanoglobule are 1 to 8 microns. If the distribution is N(5, 4), what percentage of nanoglobules meet specifications? Note that the *variance* is 4, here.

To find this, find the proportion of all nanoglobules that lie between 1 and 8:

$$\begin{aligned}\Pr(1 < x < 8) &= \Pr(\frac{1-5}{2} < \frac{x-5}{2} < \frac{8-3}{2}) \\ &= \Pr(-2 < Z < 1.5) \\ &= \Pr(Z < 1.5) - \Pr(Z < -2) \\ &= 0.9332 - 0.0227 \\ &= 0.9105\end{aligned}$$

This is done by standardizing the distribution first (see the first step), but using technology it's possible to do the math without standardizing.

There is a way to generalize the confidence interval method from much earlier. By using a new notation (z_x for quantiles), you can construct the following:

$$\bar{y} - z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}}, \bar{y} + z_{\frac{1+L}{2}} \frac{\sigma}{\sqrt{n}}$$

This will give a confidence interval for n number of samples, a $\bar{y}$ sample mean (?), a confidence level of L, and a standard deviation of σ. This assumes, of course, that the distribution of $\bar{y}$ is normal.

4.2 Central Limit Theorem

In deriving the formula for the confidence interval for a population mean, we've been assuming that the distribution of the sample mean $\bar{y}$ is normal. If the original population distribution is normal, we know that the following is correct: For a sample of size n from a $N(\mu, \sigma^2)$ population, the sampling distribution of $\bar{y}$ is $N(\mu, \frac{\sigma^2}{n})$. However, these same confidence interval formulas work well even if the population itself is not normal, due to the central limit theorem.

The theorem says that regardless of the population distribution of the quantity being measured, if the sample size is sufficiently large, the sampling distribution of the sample mean is approximately normal. This means that you can use normal distribution confidence intervals to describe a very wide range of population distributions, if you have enough data (?), at least 25 or 30 for most cases.

Example-ish In the LDL decrease study, the standard deviation was taken to be 16. If the researchers want a level 0.95 confidence interval to estimate the population mean decrease with a precision of 0.5 mg/dL, what sample size should they use?

$$n \geq (\sigma^2 + z^2_{\frac{1+L}{2}})/d^2 \approx 4000$$

4.3 Components of a Statistical Estimation Problem

The scientific goal The scientific goal is the reason for doing the experiment or study. In the above example, it was to determine if a medical treatment was effective at reducing LDL levels.

The statistical model The statistical model is the distribution of the population of measurements that are being taken. In this case, the measurements are the LDL decreases and we can assume that the population has a $N(\mu, 16^2)$ distribution.

The model parameters to be estimated At this point, examine how to achieve the scientific goal in terms of the statistical model. If you can't formulate the scientific goal in these terms, you shouldn't be doing a statistical estimation problem.

Point and interval estimates Often, the point estimator is the sample version of the model parameter to be estimated. This is true when we want to estimate the mean of an $N(\mu, \sigma^2)$ population — the estimator of μ is the sample mean $\bar{y}$. For the LDL data, the point estimate is the value of $\bar{y}$ as 21.19.

Results and interpretation These must be carefully stated. For point estimation, you are always on solid ground making a statement like "The estimate of the mean LDL reduction is 21.19 mg/dl". Generally speaking though, you should always give some indicator of the variation in the estimate, such as by giving the standard error (here: $\frac{16}{\sqrt{10}}$ or a confidence interval. When reporting a confidence interval, make a statement like "A 95% confidence interval for the mean LDL reduction is (11.27, 31.11) mg/dL", although this alone won't be enough for a quiz answer. In this case, the correct interpretation of "95% confidence" is that 95% of all possible samples will produce intervals that contain the true population mean LDL reduction.

4.4 t Distributions

If you know the population variance σ^2, calculations are fairly simple, but this is a very unrealistic expectation. If you *don't* know the population variance, the right thing to do is to use the sample standard deviation s in place of the unknown population standard deviation σ. However, replacing σ with s doesn't result in the same distribution — it results in a t distribution, also known as Student's t distribution (after the pseudonym of the person who invented it).

t distributions are actually a family of distributions that have an integer parameter ν (also written as t_ν), called degrees of freedom. The t distribution density curves look like standard normal density curves, but they are lower in the center and higher on the edge curves. In effect, this results in more variation a normal distribution.

If the original data comes from a normal distribution $N(\mu, \sigma^2)$, the standardized mean $t = \frac{\bar{y}-\mu}{s/\sqrt{n}}$ has a t distribution with $n-1$ degrees of freedom. If we let $t_{n-1,q}$ denote the q^{th} quantile of the t_{n-1} distribution, and mimic the derivation of the confidence level interval for the mean when σ is known, we get the following formula:

$$(\bar{y} - t_{n-1,\frac{1+L}{2}} \frac{s}{\sqrt{n}}, \bar{y} - t_{n+1,\frac{1+L}{2}} \frac{s}{\sqrt{n}})$$

Using the old LDL decrease data, it was previously assumed that $\sigma = 16$. Suppose that we don't actually know that value, we must instead calculate $s = 15.45$. Computing a 95% confidence interval gives you a quantile value of 2.2622, and the whole interval turns out to be $(10.14, 32.24)$. This interval is a bit wider than the previously calculated interval using a normal distribution and a value of 16 for σ, reflecting the greater uncertainty that results from estimating σ instead.

4.5 Inference for a population proportion

Using the LDL data from earlier, let p denote the proportion of the population for whom the medication will lower LDL. To estimate this value, use the proportion of the sample whose LDL decreased, which is $\hat{p} = 0.9$.

This means that $\hat{p}$ is a point estimator for p, but we want a confidence interval for this value, of 95% in this case.

Suppose we have a sample size $n = 10$ and that $\hat{p}$ is the proportion in the sample having the characteristic of interest. It can be shown that the mean and variance of $\hat{p}$ are p and $\frac{p(1-p)}{n}$, respectively. If n is large, the central limit theorem applies to $\hat{p}$, and will ensure that the distribution of $\hat{p}$ (standardize by subtracing the mean and dividing by standard error, $\sqrt{\frac{p(1-p)}{n}}$) will be approximately normal. That means that an approximate large sample level L confidence interval for p has endpoints $\hat{p} \pm z_{\frac{1+L}{2}} \sqrt{\frac{p(1-p)}{n}}$. This doesn't help much, though, because we still don't know p. Fortunately, for large samples, you can simply substitute $\hat{p}$ in for p:

$$\hat{p} \pm z_{\frac{1+L}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Note that this technically *can* result in an interval that exceeds 1.0, so in that case simply cap it off at 1.0 exactly.

Obviously, this only works well for large samples; for $n = 10$, it's not very good at all. By making an adjustment, however, this can be fixed using "fudge factors". Assume you want a level L confidence interval. Let y denote the number of items in the sample having the characteristic of interest (so $\hat{p} = \frac{y}{n}$), and compute the adjusted sample proportion:

$$\tilde{p} = \frac{y + 0.5 z^2_{\frac{1+L}{2}}}{\tilde{n}}$$

$$\tilde{n} = n + z^2_{\frac{1+L}{2}}$$

Once these values are computed, just plug them in to the original large-sample formula, and you get a confidence interval. For the LDL data, if you use these formulas and the LDL data, you could say with 95% confidence that the proportion for whom the medication will decrease LDL is between 0.574 and 1. This means that if you were to take all possible samples from a population, 95% of those samples will contain the true proportion p.

This formula has two names: the "approximate score interval", and the "Agresti-Coull confidence interval" (after the researchers that invented the idea).[1]

5 Statistical Inference for Multiple Samples

In the original LDL study, recall that each of 10 subjects obtained in a simple random sample were measured for LDL at the outset and then after 30 days on a medication. Last section's analysis focused on the average LDL decrease, but what about looking at it in a different way? These measurements were paired, so we just took a look at the difference between them and called it good.

The original data gave a point estimate of 21.9 for the population mean decrease, with a 95% confidence interval of (10.14, 32.24). What if instead we looked at it as if these two were different groups? Suppose that we take a random sample of size n_1 from population 1, which follows a $N(\mu_1, \sigma^2)$, and a second, independent sample of size n_2 from population 2, which follows a $N(\mu_2, \sigma^2)$ distribution. The sample sizes and means are not necessarily the same. The only possible difference in population distributions, however, is in their means.

We already know that the estimator of μ_1 is the sample mean of the first sample ($\bar{y}_1$), and that the estimator of μ_2 is $\bar{y}_2$. We also know the sampling distribution of $\bar{y}_1$ is $N(\mu_1, \sigma^2/n_1)$ and that the distribution of $\bar{y}_2$ is $N(\mu_2, \sigma^2/n_2)$.

The most basic method of analysis is to compare the means: $\mu_1 - \mu_2$. Common sense says to take $\bar{y}_1 - \bar{y}_2$. For this to be worth anything, we also need a confidence interval, using the information that the sampling distribution of $\bar{y}_1 - \bar{y}_2$ is:

[1] For all homework, lab assignments, and the final, use the approximate score interval for estimating a population proportion.

$$N(\mu_1 - \mu_2, \sigma^2(\frac{1}{n_1} + \frac{1}{n_2}))$$

The standardized estimator:

$$Z = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}} \approx N(0, 1)$$

...so the level L confidence interval for $\mu_1 - \mu_2$ is:

$$\bar{y}_1 - \bar{y}_2 - z_{\frac{1+L}{2}} \sqrt{\sigma^2(\frac{1}{n_1} + \frac{1}{n_2})}$$

For variance, use a pooled variance estimate (???):

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}$$

This means that you can use s_p in the standardization formula:

$$t^{(p)} = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{\sqrt{s_p^2(n_1^{-1} + n_2^{-1})}}$$

...and that gives a level L confidence interval for $\mu_1 - \mu_2$ as:

$$\bar{y}_1 - \bar{y}_2 \pm t_{n_1+n_2-2, \frac{1+L}{2}} \sqrt{s_p^2(n_1^{-1} + n_2^{-1})}$$

This is known as a pooled variance interval (???).

Example For some data, here is some calculations for two sets of data:

$$\begin{aligned}
\bar{y}_1 - \bar{y}_2 &= 108.4 - 134.9 \\
&= -26.5 \\
s_p^2 &= \frac{(10 - 1)(26.9^2) + (13 - 1)(18.4^2)}{10 + 13 - 2} \\
&= 503.6
\end{aligned}$$

Estimate standard error:

$$\begin{aligned}
&= \sqrt{503.6(\frac{1}{10} + \frac{1}{13})} \\
&= 9.44
\end{aligned}$$

Now get a 90% confidence interval:

$$\begin{aligned}
&= (-26.5 - (9.44)(1.7207), -26.5 + (9.44)(1.7207)) \\
&= (-42.7, -10.3)
\end{aligned}$$

5.1 Analysis where population variances are not equal

The most fundamental question is whether it even makes sense to compare the means, since the population variances are not equal.

The sampling distribution of $\bar{y}_1$ is $N(\mu_1, \sigma_1^2/n_1)$ and the same (but with different subscripts) for $\bar{y}_2$. The best estimator of $\mu_1 - \mu_2$ is then $\bar{y}_1 - \bar{y}_2$. To construct a confidence interval, take the sampling distribution of the estimator as $N(\mu_1 - \mu_2, \sigma_1^2/n_1 + \sigma_2^2/n_2)$.

Now standardize the estimator:

$$Z = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Of course, if you don't know the population variances, you have to estimate it, using a t distribution. Actually, scratch that, it's not a t distribution, it doesn't even have a name. The distribution can be approximated by a t distribution with v degrees of freedom, where v is the largest integer less than or equal to:

$$\frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

That's bad. Anyways, here's the final interval:

$$(\bar{y}_1 - \bar{y}_2 - t_{v,\frac{1+L}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}, \bar{y}_1 - \bar{y}_2 + t_{v,\frac{1+L}{2}} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}})$$

Example A company buys cutting blades used in its manufacturing process from two suppliers. In order to decide if there is a difference in blade life, the lifetimes of 10 blades from manufacturer 1 and 13 blades from manufacturer 2 used in the same application are compared.

The estimated standard error is:

$$\sqrt{\frac{26.9^2}{10} + \frac{18.4^2}{13}} = 9.92$$

The degrees of freedom v is calculated at 15 using a very long and annoying calculation, the formula for which can be found above. Using a 0.90 confidence interval, find $t_{15,0.95} = 1.7530$ The final interval is:

$$(-26.5 - (9.92)(1.753), -26.5 + (9.92)(1.753)) = (-43.9, -9.1)$$

This is the Satterthwaite interval thingy, as opposed to the pooled variance interval from earlier.

5.2 Comparing two population proportions

Suppose there are two populations: population 1, in which p_1 proportion has a certain characteristic, and population 2, in which a proportion p_2 has a certain possibly different characteristic.

The standard error $\hat{p}_1 - \hat{p}_2$ is:

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

For large n_1 and n_2, the central limit theorem ensures that $\hat{p}_1 - \hat{p}_2$ has approximately a normal distribution, so this has approximately a normal distribution $N(0, 1)$:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

This can be used to compute a confidence interval. If n_1 and n_2 are large, then $\hat{p}_1$ and $\hat{p}_2$ are close to p_1 and p_2, respectively. The confidence interval then has the same usual form:

$$\hat{p}_1 - \hat{p}_2 \pm z_{\frac{1+L}{2}} \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

6 Hypothesis Tests

This section will use the same LDL survey data as the previous two sections. Here's some science stuff:

- Scientific hypothesis: On average, the medication lowers LDL levels of people with high LDL.
- Statistical model: The model for the population measurements; here, the measurements are the LDL decreases and we can assume that the population has a $N(\mu, \sigma^2)$ distribution.
- Statistical hypothesis: in hypothesis testing problems, there are always two hypotheses. One is the null hypothesis H_0, and the alternative hypothesis H_a. H_a is also the scientific hypothesis, that $\mu > 0$, while H_0 is that $\mu = 0$.
- Test statistic: in all one-parameter hypothesis test settings that will be considered, the test statistic will be the estimator of the population parameter about which the inference is being made. The estimator of μ is the sample mean $\bar{y}$, which is also the test statistic here. For the LDL data, $\bar{y} = 21.19$.
- P-value: The P value can also be thought of as a plausibility value. The basic question is whether the data backs up the null hypothesis or not. To compute it, first assume that the null hypothesis is true. The p-value is then the proportion of all samples from this population for which the test statistic will give as much or more evidence against H_0 and in favor of H_a as does the observed test statistic value. In this case, any $\bar{y}$ value larger than the observed value of $\bar{y} = 21.19$ will provide as much or more evidence against H_0 and in favor of H_a as does the observed test statistic value. Thus, the p-value is $\Pr(\bar{y} \geq 21.19)$, where $\Pr(expression)$ represents the proportion of all samples for which *expression* is true, computed under the assumption that H_0 is true.

Time to calculate a real p value. Here we'll assume that we know the population variance $\sigma = 16$. To calculate the p-value, standardize the test statistic by subtracting its mean (which is $\mu = 0$ here, since we're assuming the null hypothesis is true) and dividing by the standard error $\sigma/\sqrt{n}$. If H_0 is true, the result will have a $N(0,1)$ distribution.

For this data:

$$\begin{aligned}
\frac{16}{\sqrt{10}} &= 5.060 \\
\Pr_0(\bar{y} \geq 21.19) &= \Pr_0(\frac{\bar{y} - 0}{\sigma/\sqrt{10}} \geq \frac{21.19}{\frac{\sigma}{\sqrt{10}}} \\
&= \Pr_0(N(0,1) \geq 4.188) \\
&= 10^{-5}
\end{aligned}$$

This means that $\bar{y}$ is pretty far out there, specifically on a Z value of 4.188. This is not consistent with the null hypothesis, so the null hypothesis must be rejected.

Example 2 In the case where you don't know the population variance σ^2, you must estimate it using the sample variance s^2. The p-value is then $\Pr(t_{n-1} \geq t*)$. Here's an example:

$$\begin{aligned}
s &= 15.45 \\
t* &= \frac{21.19 - 0}{15.45/\sqrt{10}}
\end{aligned}$$

(i.e. standardize and divide by the standard error). The P value can then be calculated using a t distribution with 9 degrees of freedom (since there were 10 data points), and the final p value is $9.4 \cdot 10^{-4}$.

If the p-value is small enough, it indicates that, relative to H_a, the data is not consistent with the assumption that H_0 is true (that is, that the data is not plausible). That means that the proper action is to reject H_0 in favor of H_a. Whether the p value is small enough depends on the type of study, but here are some guidelines:

- 0.100 — borderline
- 0.050 — reasonably strong (traditional value)
- 0.025 — strong
- 0.010 — very strong

6.1 Two sided tests

Suppose in the LDL reduction problem that the researchers wanted to see if there was no mean change in the LDL levels as opposed to the medication making some difference, either good or bad. The appropriate hypotheses would then be $H_0 : \mu = 0$ and $H_a : \mu \neq 0$.

To compute the p value of a two sided test, first compute the standardized test statistic t and its observed value $t*$:

$$\begin{aligned} t &= \frac{\bar{y} - \mu_0}{S/\sqrt{n}} \\ t* &= \frac{21.19 - 0}{4.886} \\ &= 4.337 \end{aligned}$$

Because the test is two sided and symmetrical, the p value can be computed as $2\Pr(t_9 \geq |t*|)$. In this case, that means $2\Pr(t_9 \geq 4.337) = 18.8 \cdot 10^{-4}$. There's a useful formula for this too: $2\Pr(t_9 \geq |t*|) = 2min(\Pr(t_9 \leq t*), \Pr(t_9 \geq t*))$ For this LDL value data, that still shows that the results are very significant.

6.2 Philosophy of Hypothesis Testing

Statistical significance is in essence a measure of our ability to detect a difference. Therefore, it's a reflection of not only the size of the difference but also the amount of data (evidence) available. Suppose that for the previous LDL study, the same study was run with the same results (mean, standard deviation, etc.), but only three subjects. Running those calculations again yields a much higher standard error and a different p value, of 0.0703. This result is much less significant and ultimately ends in a different conclusion (that we cannot reject H_0). The practical significance is the same here, but statistically we can't claim to observe the same difference – it's indistinguishable from background noise.

There are a number of things to be wary of:

- Using the same data to suggest and confirm hypotheses (the difference between an exploratory study and a confirmatory study) isn't a good idea.
- Doing lots of tests results in lots of false positives.
- Lack of significance is not failure. The result *is* important, but it isn't failure.

6.3 Testing a population proportion

Let p be the proportion of a population that has some given characteristic. Hypotheses about p are often tested by taking a random sample and observing the number of sample units that show the characteristic.

Using the LDL data (again, this study never dies...), researchers want to test to see if the medication results in reductions in LDL levels in more than half of al people with high LDL. That means that $H_0 : p = 0.5$ and $H_a : p > 0.5$. (note: this class will not show the mechanism for doing the more proper H_0 test where $H_0 \leq 0.5$). Recall that the estimator for p is $\hat{p} = y/n$, where y is the number in the sample showing the characteristic and n is the sample size. The test statistic used is y, which gives the same results as if $\hat{p}$ was used, for some reason.

The p-value will be given by $\Pr_0(y \geq y*)$ where $y*$ is the number in the sampel who have the characteristic, and $\Pr_0(y \geq y*)$ is the proportion under H_0 (that is, when $p = 0.5$) of all size n samples from the population for which y will be at least as large as $y*$. For this data set, $y* = 9$.

The proportion of all random samples in which exactly y have the characteristic is given by this formula:

$$\Pr(y = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

... for $k = 0, 1, 2, \ldots n$. This is called a binomial distribution and is the sampling distribution of y, and is written as $b(n, p)$. For this example, the distribution is $b(10, 0.5)$. When H_0 is true, the proportion of all samples of size 10 in which exactly k have lower LDL is given by:

$$\Pr_0(y = k) = \binom{10}{k} 0.5^k (1-0.5)^{10-k}$$

The proportion itself is calculated by:

$$\begin{aligned} \Pr_0(y \geq 9) &= \sum_{k=9}^{10} \Pr_0(y = k) \\ &= 0.0107 \end{aligned}$$

This can be considered to be strong evidence against H_0 and in favor of H_a, so we could justifiably conclude that the medication will decrease LDL levels for more than half of the population.

For a random sample of n observations from a large population with a proportion p having a characteristic, where $y*$ and $n - y*$ are large (at least 10), the observed value of the standardized test statistic (with continuity correction of 0.5) is:

$$z*_l = \frac{y* - np_0 + 0.5}{\sqrt{np_0(1-p_0)}}, z*_u = \frac{y* - np_0 - 0.5}{\sqrt{np_0(1-p_0)}}$$

This has something to do with something called a large sample test.

This does something something something standardization, with the continuity correction to make approximation by the normal curve (seriously, what?!) more accurate, something to do with binomials only taking on integer values and trying to bring values to between integers. P values are then:

$$\begin{aligned} p = p_0 : p < p_0 : p_- &= \Pr(N(0,1) \leq z*_l) \\ p = p_0 : p > p_0 : p_+ &= \Pr(N(0,1) \geq z*_u) \\ p = p_0 : p \neq p_0 : p_\pm &= \Pr 2\min(p_+, p_-) \end{aligned}$$

6.4 Comparing two population means: paired data

In the LDL reduction study, each of the 10 subjects were measured twice — one before treatment, one after treatment. The analysis previously relied on testing the mean reduction in LDL. Looked at another way, the mean reduction is the difference of two population means. In the hypothesis setting, this just means that we change the interpretation of the hypotheses:

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_a : \mu_1 - \mu_2 > 0$$

Assume a random sample of n paired observations $(y_{1,i}, y_{2,i})$ where $i = 1 \dots n$. The mean of $y_{1,i}$ is μ_1 and the same for μ_2. The observed value of the test statistic is:

$$t* = \frac{\bar{d}* - \delta_0}{s_d/\sqrt{n}}$$

... where δ_0 defines the null hypothesis. The P value table thingy looks about the same, except with a t_{n-1} distribution.

6.5 Comparing two population means: independent populations

Suppose we take a random sample of size n_1 from population 1, which follows a $N(\mu_1, \sigma^2)$ distribution, and do the same thing for a sample from population 2. For now, the variance will remain the same. We already know the estimators of μ_1 and μ_2, and the sampling distribution should be $N(\mu, \sigma^2/\sqrt{n})$. The pooled variance estimate looks about the same:

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2)}{n_1 + n_2 - 2}$$

The null hypothesis is that $\mu_1 - \mu_2 = \delta_0$ where δ_0 is some postulated number, usually zero.

$$t^{(p)} = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{\sqrt{s_p^2(n_1^{-1} + n_2^{-1}}}$$

Assuming $t^{(p)*}$ is the observed value of $t^{(p)}$, the hypothesis tests are the same as that 3-line table thingy from a few pages ago. Note that there is a t distribution involved here somewhere.

Once you have a value for $t^{(p)}$, you find the corresponding t distribution value for a distribution with $n_1 + n_2 - 2$ degrees of freedom.

If the population variances are not equal, you must use that annoying Satterthwaite distribution thing:

$$t^{(ap)} = \frac{\bar{y}_1 - \bar{y}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

... I am NOT typing out that hell fraction from earlier. It's the thing that calculates ν. Anyways,

$$t^{(ap)*} = \frac{\bar{y}_1 - \bar{y}_2 - \delta_0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Just use a t distribution with ν (round down) degrees of freedom and use the same hypothesis test table that I still won't duplicate because it's annoying to write.

6.6 Comparing two population proportions

$\hat{p}_1$ and $\hat{p}_2$ are the population proportions for our two different population proportions (assume that all the other variables are still the same here). A point estimator of $p_1 - p_2$ is then just $\hat{p}_1 - \hat{p}_2$. The standard error of this is:

$$\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

If n_1 and n_2 are large, $\hat{p}_1 - \hat{p}_2$ are approximately normal (due to the central limit theorem), so this thing has about an $N(0,1)$ distribution:

$$\frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}}$$

The same thing from above about δ_0 applies here. Since this is a large sample test and thus relies on the central limit theorem, a good rule of thumb is that $y_i \geq 10$ and $n_i - y_1 \geq 10$ for $i = 1, 2$.

Suppose H_0 is that $p_1 - p_2 = 0$ (i.e. $\delta_0 = 0$). That means that $p = p_1 = p_2$, and if H_0 is true, the variance of $\hat{p}_1$ is $\frac{p(1-p)}{n_1}$ and the same-ish for $\hat{p}_2$. That means that the standard error of $\hat{p}_1 - \hat{p}_2$ is:

$$\sqrt{p(1-p)(n_1^{-1} + n_2^{-1}}$$

We don't know p, though, so estimate it using data from both populations:

$$\hat{p} = \frac{y_1 + y_2}{n_1 + n_2}$$

The standardized test statistic is then:

$$Z_0 = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{p}(1-p)(n_1^{-1} + n_2^{-1})}$$

The observed value of Z_0 is denoted as z_0^*. This can be plugged into the constantly-referenced hypothesis table to find an actual p value.

Now suppose that $\delta_0 \neq 0 \ldots$

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

Keep in mind, the rule of thumb for central limit theorem sample sizes still applies here.

6.7 Fixed Significance Level Tests

An alternative way to conduct a test at a fixed level of significance is to determine which values of the test statistic will lead to rejection of H_0 in favor of H_a.

Using the LDL data again (of *course*), specify the null hypothesis to be that the mean is zero, and the alternative hypothesis to be that the mean is greater than zero. Set the significance level α to 0.05. The standardized test statistic is then:

$$t\frac{\bar{y} - \mu_u}{s/\sqrt{n}} = \frac{\bar{y} - 0}{15.45/\sqrt{10}}$$

None of this should be new, but what comes next should be. Now we find the critical region of the test, or the set of values of the (standardized) test statistic for which we will reject H_0 in favor of H_a. H_a tells us that the critical region has the form of $(t_{9,0.95}, \infty)$. There is no p value computed, we simply decide that if the test statistic does not fall into this region, we accept H_0. Here, that means that H_0 will be rejected if and only if the observed value of t is greater than or equal to 1.8331 (based on the t-table).

Using the LDL data, the test statistic is 4.337, which is much bigger than 1.8331, and thus lies within the critical region. Because of this, we must reject the null hypothesis. This is the same result as using p values, it just uses a different approach.

This shows a relationship between hypothesis tests and confidence intervals. Consider the following two inference procedures for a population mean μ, conducted on the same set of data. You could use a two-sided α test of $H_0(\mu = \mu_0)$, verses a level L confidence interval. The test will reject H_0 in favor of H_a if and only if μ_0 lies outside the confidence interval, so you can just use the confidence interval to conduct the test. Similarly, using a level $1 - \alpha$ confidence interval for μ consists of all μ_0 values for which the test does not reject H_0 in favor of H_a. This is called "inverting the test" [to obtain a confidence interval].

Of note is that you can create one-sided confidence intervals (as of yet all of these confidence intervals have been two sided) by inverting tests of one sided alternative hypotheses. For example, if you wish to test $H_0 : \mu = \mu_0, H_a : \mu > \mu_0$, a level α test rejects H_0 in favor of H_a if:

$$t_{n-1,1-\alpha} \leq \frac{\bar{y} - \mu_0}{s/\sqrt{n}} < \infty$$

With some algebra, this is the same thing as:

$$-\infty < \mu_0 \leq \bar{y} - \frac{s}{\sqrt{n}} t_{n-1,1-\alpha}$$

6.8 Power of a Test

In a fixed significance level test, power is the proportion of all samples for which H_0 will be rejected in favor of H_a. Power will vary for different values of the parameter being tested, so it is written as a function of that parameter.

Suppose you take a sample of size n taken from a $N(\mu, 25)$ population. We want to test $H_0 : \mu = 10$ and $H_a : \mu < 10$ at a 0.05 level of significance using a fixed significance level test. The form of H_a tells us that small values of $\bar{y}$ should lead to rejection of H_0, which means small values of the standardized test statistic should lead to rejection. After a little bit of math, the rejection/critical region becomes $\bar{y} \leq 10 - (1.645)(5)/\sqrt{n}$. To compute the power of this test, we need to evaluate the proportion of all samples which lead to rejection when the true population mean is μ. Write this out as:

$$\begin{aligned} \Pi(\mu) = \Pr_{\mu}(reject H_0) &= \Pr_{\mu}(\bar{y} \leq 10 - (1.645)(5)/\sqrt{n}) \\ &= \Pr(z \leq \sqrt{n}(10 - \mu)/5 - 1.645) \end{aligned}$$

For any value of μ, $\Pi(\mu)$ can be computed using computers, a calculator, or a table of the normal distribution. For instance, if $n = 16$ and $\mu = 7$, we get $\Pr(z \leq 0.755) = 0.775$.

Since the power curve depends on sample size n, the power function can be used to specify a sample size. If the researcher specifies a significance level for the test and a desired power at a specified value of the parameter being tested, then using the formulas given, a value for n can be determined. This is often doing computers.