Review of Math 161A material

The Central Limit Theorem

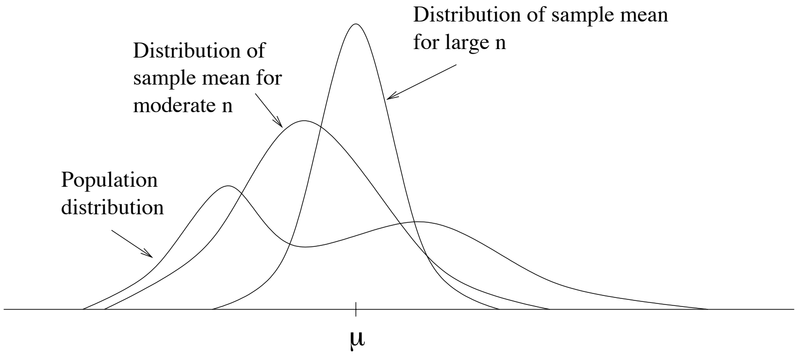
Remark: The Central Limit Theorem (CLT) is one of the most important results in probability theory. It was first stated by Abraham DeMoivre in 1733 and generalized by Pierre Laplace in 1812. Because of that, it is also sometimes called the De Moivre - Laplace Theorem. It took statisticians over 150 years (until 1901) to prove the Central Limit Theorem in its most general form.

Situation: Consider an experiment where a certain variable is measured repeatedly many times. Let X_i be the result of the i^{th} measurement. The X_i's may have some discrete or continuous (not necessarily known) distribution. What can we say about the sum or the average of the X_i's?

Fact: If the X_i's have normal distributions, then the sum or average of the X_i's is exactly normally distributed.

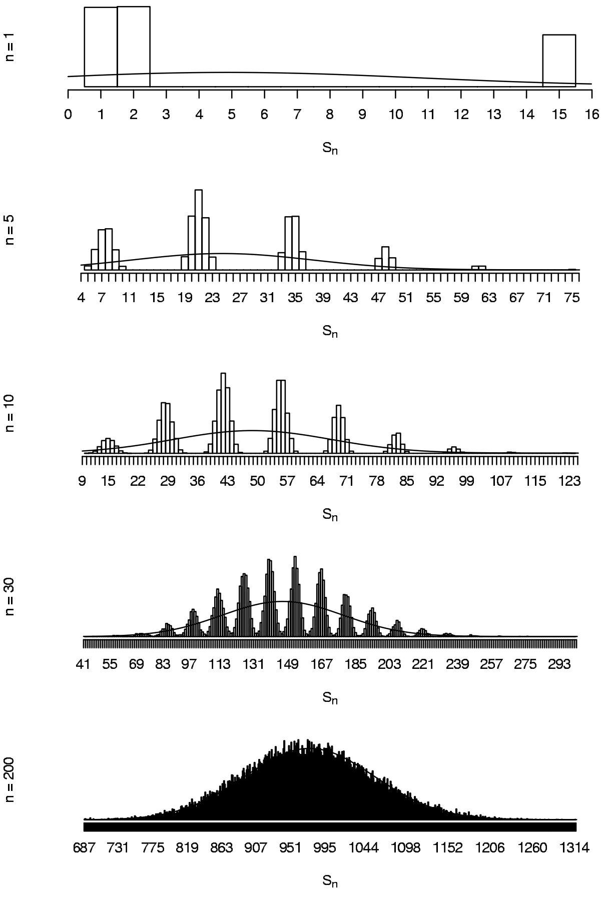
Amazing Fact: If a measurement is repeated often, then the sum or the average of any kind of random variable (discrete or continuous, not necessarily normal) is approximately normally distributed.

The larger n is, the more closely will the PDF of the sum or average resemble a Normal PDF.


Distribution of sample mean for large n
Distribution of sample mean for moderate n
Population distribution
μ

In General: In some textbooks, values of n are considered large if they are $n > 30$. Other textbooks use $n > 20$ or $n > 40$. In reality, how large n should be depends on how "non-normal" the distribution of the random variables is that are summed or averaged. The more "non-normal" the distribution is, the larger n should be for the Central Limit Theorem to kick in.

Shown below is the PMF of the sum of n discrete X_i for various n-values. Note, that the distribution of a single X_i (first graph for $n = 1$) is quite non-normal.



Theorem: Central Limit Theorem

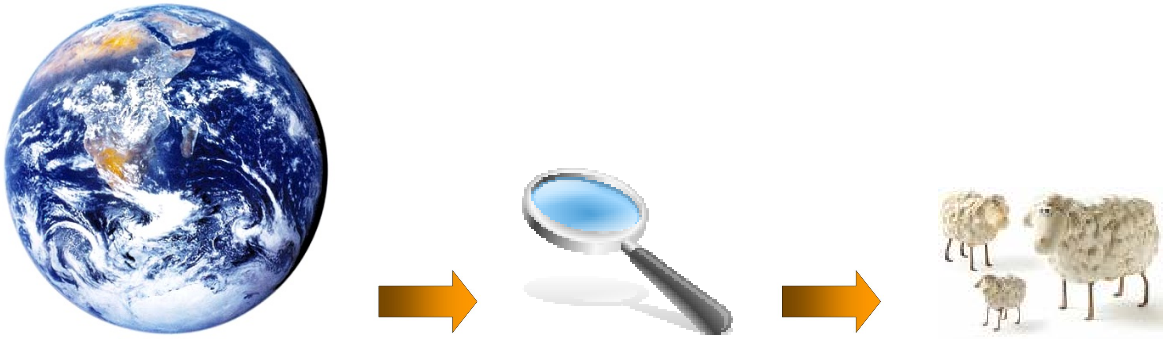
Let $X_1, \ldots, X_n$ be independent random variables with $E(X_i) = \mu$ and $\text{Var}(X_i) = \sigma^2$ for $i = 1, \ldots, n$. Then for large n ($n > 30$), the sum and the average of the X_i's has approximately a Normal distribution

$$\boxed{\begin{array}{lrcl} \text{Average:} & \bar{X} = \frac{1}{n}\sum\limits_{i=1}^{n} X_i & \sim & \text{Normal}(\mu, \frac{\sigma^2}{n}) \\ \text{Sum:} & \sum\limits_{i=1}^{n} X_i & \sim & \text{Normal}(n\mu, n\sigma^2) \end{array}}$$

Example: Let $X_1, \ldots, X_{40}$ be independent exponential random variables with $\lambda = 1/2$. Find $P(\bar{X} \leq 1)$.

Point Estimation

Recall: The goal of statistics is to use data to draw conclusions about populations. In particular, we are interested in one or more specific characteristics of the population. These (unknown) population parameters are usually denoted by Greek letters. Since it is often not feasible to investigate every member of the population, a random sample is taken instead.



Population with unknown parameter μ

Take a sample $X_1, \dots, X_n$

Compute a statistic $\hat{\mu} = \bar{x}$

Definition: A point estimate of a parameter θ is a number that can be regarded as a sensible value for θ. The point estimate is obtained by choosing a suitable statistic and computing the value of the statistic for a random sample taken from the population.

Example: For each of the following scenarios suggest a sensible statistic that would lead to a point estimate for the population parameter.

(a) Let p be the proportion of humans in sub-Saharan Africa who carry the HIV virus. Suppose you have obtained blood samples from n randomly chosen individuals.

(b) Let μ be the average yield (in tons) produced from one acre of cotton plants. You randomly select n cotton farmers who report the yields on one acre parcels of land.

(c) Let σ be the standard deviation of annual rate of return of a particular investment. You have data on the annual rate of return for the past n years.

Popular point estimates: Some often used point estimates include

Sample proportion: $\hat{p} = \frac{\text{\# of individuals in sample which exhibit a trait}}{n}$

Sample mean: $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$

Sample variance: $s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$

Note: A point estimate $\hat{\theta}$ for θ is a random variable. It depends on the choice of the statistic and on the selection of the sample. Hence, we cannot always expect that our "best guess" $\hat{\theta}$ will be exactly equal to the true population parameter.

$$\hat{\theta} = \theta + \text{ error of estimation}$$

Unbiased Estimators

DEFINITION: A point estimate $\hat{\theta}$ is said to be UNBIASED, if

$$E(\hat{\theta}) = \theta$$

This means that if we keep obtaining "best guesses" $\hat{\theta}$ for θ, their average will converge to the true answer θ. If a point estimate $\hat{\theta}$ is not unbiased, then the difference $E(\hat{\theta}) - \theta$ is called the BIAS of $\hat{\theta}$.

Standard Error

DEFINITION: The STANDARD ERROR of an estimator $\hat{\theta}$ is its standard deviation $\sigma_{\hat{\theta}} = \sqrt{V(\hat{\theta})}$. If the standard error itself involves unknown parameters that can be estimated the substitution of those estimates into $\sigma_{\hat{\theta}}$ yields the ESTIMATED STANDARD ERROR $s_{\hat{\theta}}$.

Example: The Central Limit Theorem states that the distribution of the sample mean $\bar{x}$ of independent and identically distributed random variables $X_1, \ldots, X_n$ is

$$\bar{X} \sim \text{ Normal}\left(\mu, \frac{\sigma^2}{n}\right)$$

where $\mu = E(X_i)$ and $\sigma^2 = V(X_i)$. What can we say about the point estimate $\bar{x}$ for μ?

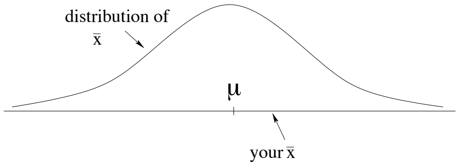
Confidence Intervals

Recall: A point estimate is a "best guess" function that computes an estimate for a population parameter θ based on sample data. The value of the point estimate depends on the function used and the sample data. Since the sample is selected at random, the point estimate is a random variable. For each sample taken, we get to see only the result - one number. This number by itself does not provide information on the precision and reliability of the estimate.

Example: 100 people at a fair guess the number of Jelly Beans in a large glass jar. You, cleverly, record their guesses and compute the average to hand that in as "your" guess in hopes of winning the grand prize of a life-time jelly bean supply.

Identify the population parameter and the point estimate(s) in this example. What do you know about the distribution of the point estimate(s)?

Usually, point estimate functions are chosen, because we hope that their computed values for random samples are close to the true (unknown) population parameter. But how close?



If you know the distribution of $\bar{x}$, you can compute the probability that $\bar{x}$ will fall into any specified interval. Especially, you can compute the probability that $\bar{x}$ will be no further then c from μ.

Example: (cont.) Suppose that there are actually 1250 beans in the jelly bean jar. The guesses of individual people are normally distributed with mean $\mu = 1250$ and standard deviation $\sigma = 70$. You collect data on the guesses of 100 independent people and compute the average $\bar{x}$.

(a) What is the probability that a single persons guess will be within 10 jelly beans of the true answer?

(b) What is the probability that your *average* will be within 10 jelly beans of the true answer?

In General: Assume that $x_1, \ldots, x_n$ is a random sample from a Normal distribution with (unknown) mean μ and (known) variance σ^2. According to the Central Limit Theorem, $\bar{X}$ has a Normal distribution with mean μ and variance σ^2/n. Hence,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{ Normal}(0,1)$$

$$\Leftrightarrow P(-1.96 \leq \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \leq 1.96) = 0.95$$

$$\Leftrightarrow P\left(\qquad\qquad\qquad\qquad \leq \mu \leq \qquad\qquad\qquad\qquad \right) = 0.95$$

Main Idea: If $\bar{X}$ is close to μ with 95% probability, then μ is also close to $\bar{X}$ with 95% probability. Since we want to estimate μ from our best guess $\bar{X}$, we can now provide a Confidence Interval for μ.

Definition: After observing a random sample $(x_1, \ldots, x_n)$ of size n and computing the sample mean $\bar{x}$, a 95% confidence interval for μ is

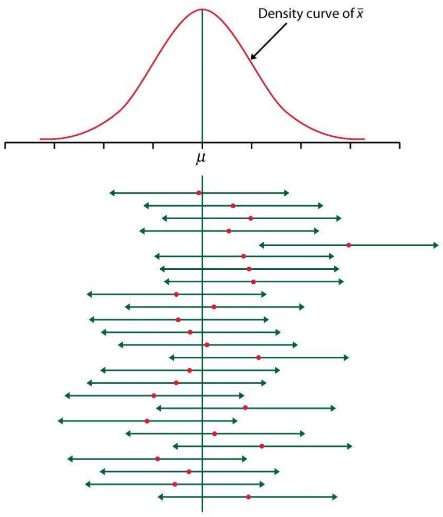
$$\text{CI}_\mu = \left[\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right]$$

Interpreting Confidence Intervals

Caution: Assume that you collect data $(x_1, \ldots, x_n)$ and then compute a 95% confidence interval for μ based on $\bar{x}$. It is tempting (but wrong) to conclude that your confidence interval contains the true μ with probability 0.95. Since the true μ is a fixed number, it will be contained in *any* interval with probability either 0 or 1. The Confidence level (here 95%) rather refers to the generation of the random variables $x_1, \ldots, x_n$ or the sampling process.

Suppose the fair runs for 20 days and you go each day and observe 100 different people, compute the average of their counts and hand that in as your "best jelly bean guess of the day". On some days you will come close to the true number on the jar (i.e., μ will be in your confidence interval) but on other days you will not be so close.

On average, for 95% *of samples* your confidence interval will contain the true population parameter μ. If you only play on one day, you do not know whether this day's confidence interval does or does not contain μ, but you are 95% confident that it does.



Confidence Intervals for Large Samples

If the sample size n is large then the sample mean has approximately a Normal distribution regardless of the population distribution (according to the Central Limit Theorem). However, in most practical applications the population variance σ^2 is not known. It is still possible to obtain confidence intervals for the mean, by first estimating $s = \hat{\sigma}$ and then using the estimate s in the confidence interval computation.

PROPOSITION: If n is sufficiently large ($n > 40$), then the standardized variable

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \sim \text{ Normal}(0, 1)$$

has approximately a standard normal distribution. The n needs to be a little bit larger here than the ($n > 30$) rule we used for the CLT because of the additional variation introduced through the estimation of s.

PROPOSITION: Let $x_1, \ldots, x_n$ be a large random sample from some population with mean μ. Then

$$\text{CI} = \left[\bar{x} - z_{\alpha/2}\frac{s}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{s}{\sqrt{n}}\right]$$

is approximately a $100(1-\alpha)\%$ confidence interval for μ.

Example: Suppose the instructor of a large statistics class tries to write an exam of medium difficulty (mean score = 75). Fifty students take the exam and their scores are:

$$88 \quad 93 \quad 74 \quad 81 \quad \cdots \quad 98$$

The mean of these exam scores is 82 with a standard deviation of 18.

(a) Compute a 95% confidence interval for the degree of difficulty of the exam.

(b) In light of this information, did the exam turn out to be easier, harder or about as hard as the instructor had intended it to be?

Confidence Interval for a Population Proportion

Let p denote the (true, but possibly unknown) proportion of "successes" in a population. A random sample of n individuals is selected from the population and let X denote the number of successes in the sample. Then X has approximately a Binomial distribution if the population size is large compared to the sample size. And furthermore, if the sample size n is large as well, then X has approximately a Normal distribution.

The most commonly used point estimate for p is $\hat{p} = X/n$ the sample proportion of successes.

Proposition: A $100(1-\alpha)\%$ confidence interval for the population proportion p is

$$\left[\frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + (z^2_{\alpha/2})/n}, \frac{\hat{p} + \frac{z^2_{\alpha/2}}{2n} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n} + \frac{z^2_{\alpha/2}}{4n^2}}}{1 + (z^2_{\alpha/2})/n}\right]$$

This looks terribly complicated! But if the sample size n is large, then the above formula is approximately equal to

$$\mathrm{CI}_p^{(1-\alpha)100\%} = \left[\hat{p} - z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}, \hat{p} + z_{\alpha/2}\sqrt{\frac{\hat{p}\hat{q}}{n}}\right]$$

here $\hat{p}$ is your sample proportion of "successes" and $\hat{q} = 1 - \hat{p}$ is the sample proportion of "failures".

The simple confidence interval above can be used in cases where the sample size n is large enough so that $n\hat{p} \geq 10$ and $n\hat{q} \geq 10$.

Example: In a final pre-election poll before the 2012 presidential election 49% of 789 randomly chosen registered voters in Ohio said that they favored Barack Obama (Romney 49%, Other 1%, Not Sure 1%).

Compute an approximate 95% confidence intervals for the percentage of voters in Ohio who voted for Barack Obama.

Confidence Intervals for Normal Populations

If we know the population distribution to be Normal, then the sample size may be as small as $n = 2$ for the following to hold:

Proposition: If the population is Normal, and $X_1, \ldots, X_n$ is a random sample from the population, then

$$\bar{X} \sim \text{ Normal}(\mu, \sigma^2/n) \quad \text{or} \quad Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{ Normal}(0,1)$$

If σ is unknown and has to be replaced by the estimated standard deviation s, then the distribution changes.

Proposition: Let $\bar{X}$ be the mean of a random sample of size n from a Normal population with mean μ. Then the random variable

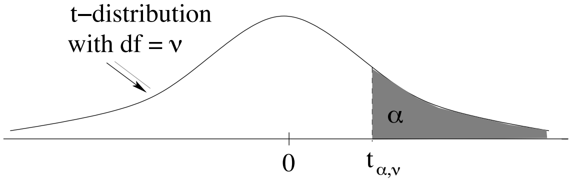
$$T = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

has a t-distribution with $n - 1$ degrees of freedom.

Remark: t-distributions look very similar to Normal distributions. They are symmetric, centered at zero and bell shaped. However, they have slightly "thicker" tails than Normal distributions. Values of t-distributions are available in tables (in the back of your textbook) or by using your calculator, Excel (commands T.INV, T.DIST), or statistical software.

Notation: Let $t_{\alpha,\nu}$ be the number on the x-axis for which the area under the t-distribution curve with df = ν to the right of $t_{\alpha,\nu}$ is equal to α. $t_{\alpha,\nu}$ is called a t critical value.

Example: Use the table in the back of the book to find $t_{0.05,4}$. Find the same number with the Excel command "=T.INV.2T(0.1,4)".



Proposition: Let $\bar{x}$ and s be the sample mean and sample standard deviation computed from a random sample of size n from a Normal population with mean μ. Then a $100(1-\alpha)\%$ confidence interval for μ is

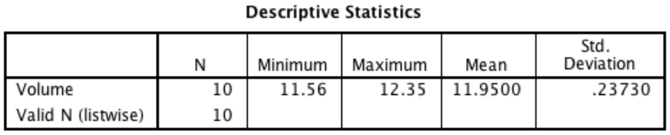
$$\left[\bar{x} - t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}, \bar{x} + t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}\right]$$

Example: A machine is used to fill cans of soft drinks. The machine should be adjusted so that the amount of soft drink in each can is approximately 12 fl oz. Of course it is not possible to fill each can with exactly the same amount, so that the amount of drink per can may be regarded as a Normal random variable with mean 12 and unknown variance σ^2.

We want to check whether the machine is adjusted correctly and randomly select $n = 10$ cans from the production line and measure precisely the amount of drink in each can:

$$12.08, 11.84, 11.97, \ldots, 12.11$$

Suppose the average $\bar{x}$ of the measurements is 11.95 with sample standard deviation $s = 0.2373$.

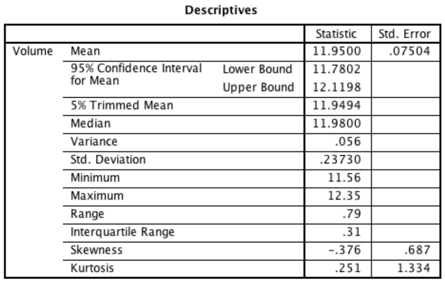


Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
Volume	10	11.56	12.35	11.9500	.23730
Valid N (listwise)	10				

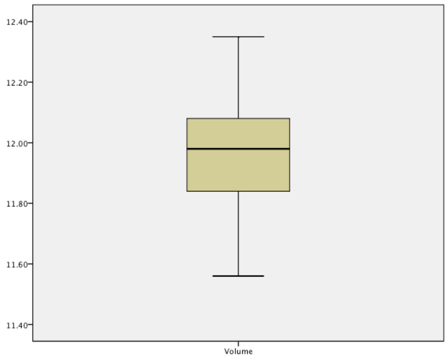
(a) Compute a 95% confidence interval for μ the average amount of drink with which the cans get filled.

To achieve the same result in SPSS, click on Analyze → Descritive Statistics → Explore... → OK.



Descriptives

			Statistic	Std. Error
Volume	Mean		11.9500	.07504
	95% Confidence Interval for Mean	Lower Bound	11.7802	
		Upper Bound	12.1198	
	5% Trimmed Mean		11.9494	
	Median		11.9800	
	Variance		.056	
	Std. Deviation		.23730	
	Minimum		11.56	
	Maximum		12.35	
	Range		.79	
	Interquartile Range		.31	
	Skewness		-.376	.687
	Kurtosis		.251	1.334



(b) Based on your answer above, do you think it is necessary to adjust the machine?

Hypothesis Testing

Recall: So far, we have discussed point estimates - finding a "best guess" for an unknown population parameter and confidence intervals - plausible values for the unknown population parameter. The objective of a statistical hypothesis test is to decide which one of two contradictory claims about the population parameter is correct.

Definition: A statistical hypothesis is a claim about a population parameter (such as the mean μ or a proportion p). The null hypothesis, denoted by H_0 is the claim that is initially assumed to be true. The alternative hypothesis, denoted by H_a is an assertion that is contradictory to H_0.

If the observed data is plausible (has high probability) under the null hypothesis assumption, we will accept this claim as true. If the observed data has very low probability under the null hypothesis and much higher probability under the alternative hypothesis, we will reject the null hypothesis in favor of the alternative.

Testing procedure: A statistical hypothesis test consists of several components.

1. The null hypothesis statement and the alternative statement. The null hypothesis should always be phrased as an equality (e.g., $H_0 : \mu = 0$ or $H_0 : p = 0.5$). The alternative can be phrased as an equality (e.g., $H_a : \mu = 3$) or an inequality (e.g., $H_a : \mu \neq 0$ or $H_a : p > 0.5$).

2. A test statistic. This is a function whose value can be computed from the sample data. We have to know the (theoretical) distribution of the function if the null hypothesis H_0 is true. The decision whether to accept or reject H_0 is based on the value of the test statistic computed from the data.

3. A rejection region - the set of all test statistic values for which the null hypothesis H_0 will be rejected.

Errors: There are two possible errors that can be made in hypothesis testing:

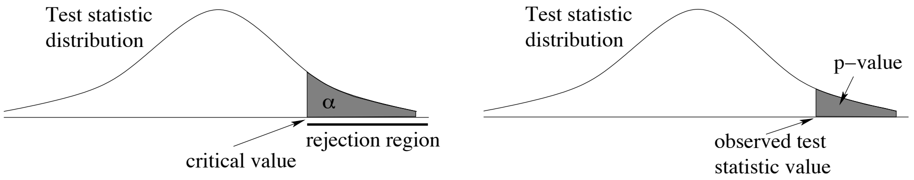
- Rejecting the null hypothesis H_0 when it is true (type I).
- Accepting the null hypothesis H_0 when it is false (type II).

Ideally, one would want to keep the probabilities of both these errors as small as possible. However, the error probabilities are related and if one error probability is made smaller the other one usually will increase. The choice of rejection region determines the probabilities of both a type I and type II error.

Definition: The probability of a type I error α is called the significance level of the test. The probability of a type II error is usually denoted by β. The quantity $1 - \beta$ represents the test's ability to correctly reject a false null hypothesis and is called the power of the test.

P-Values

One way to report the results of a hypothesis test is to say whether or not the test statistic value fell into the rejection region and subsequently, whether or not the null hypothesis was rejected at a specified level of significance α. This yes/no decision does not convey any information about how soundly the null hypothesis was rejected. *Where* in the rejection region did the observed test statistic value fall?


Test statistic distribution
α
rejection region
critical value
Test statistic distribution
p–value
observed test statistic value

DEFINITION: Suppose the null hypothesis H_0 is, in fact, true. The p-value is the probability to observe a test statistic value at least as contradictory to H_0 as the computed value by random chance due to the selection of the sample.

If a p-value is smaller than the significance level α, then the corresponding value of the test statistic falls into the rejection region and the null hypothesis will be rejected.

$$p \leq \alpha \quad \Rightarrow \quad \text{reject } H_0$$

If the p-value is large, then it is quite likely to see data such as the observed by random chance if the null hypothesis were true and H_0 will not be rejected.

$$p > \alpha \quad \Rightarrow \quad \text{fail to reject } H_0$$

IN GENERAL: Follow this procedure whenever you conduct a statistical hypothesis test:

1. Identify the parameter of interest in the problem (e.g., mean μ or proportion p, etc.)
2. Formulate the null hypothesis H_0 and the alternative hypothesis H_a.
3. Select a test statistic and compute the test statistic value for the sample data.
4. (a) EITHER: Find the rejection region for your type of alternative and level of α. Determine whether or not your test statistic value falls into the rejection region.

   (b) OR: (better!) compute the p-value for your observed test statistic value. Compare the p-value to α. Reject H_0 if $p < \alpha$, accept H_0 if $p \geq \alpha$.
5. Draw a conclusion and decide whether or not to reject H_0. Your conclusion should always be formulated as a sentence (not a formula) and be worded in the context of the original example.

Tests for a Population Mean

If a sample is large, then the population standard deviation σ may be replaced by its point estimate s without changing the distribution of the sample mean:

$$\text{Test statistic:} \qquad Z = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim \text{ Normal}(0,1)$$

Since the Normal distribution is also sometimes referred to as the z-distribution, this kind of test is called a z-test for a single population mean. If the sample size n is not large but the distribution of X is Normal, then replacing the population variance σ with the sample variance s changes the distribution of the sample mean from Normal to the t-distribution.

$$\text{Test statistic:} \qquad T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}} \sim t(df = n-1)$$

The testing procedure in this case is very similar to the z-test procedure, except that the rejection region (or p-value computation) now depends on the t-distribution instead of the Normal distribution.

Example: Past data have shown that if parking meters in a very small town are emptied every 14 days, the coin collectors will be about 70% full. The collection agency has planned the visits this way because if the meters are full they become unusable, but emptying the meters too often increases employment costs. During the last visit five randomly selected meters were 50%, 40%, 70%, 75%, and 45% full, respectively. Do you think the frequency of visits should be changed?

Test for a Population Proportion

Let p denote the proportion of individuals in a population who possess a certain characteristic (successes). A random sample of size n is selected from the population and the proportion $\hat{p}$ of successes in the sample is observed. We want to use this quantity to decide whether to accept or reject a statement about the population parameter p.

Large Sample Tests

If the sample size n is large ($np_0 \geq 10$ and $n(1-p_0) \geq 10$), then the test statistic for the hypothesis test

$$H_0 : p = p_0$$

has approximately a standard Normal distribution

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}} \sim \text{ Normal}(0,1).$$

Example: Prevnar is a vaccine for meningitis usually given to infants. In a clinical trial, Prevnar was given to 710 children, of whom 72 experienced a loss of appetite. Competing medications cause about 13.5 percent of children to experience a loss of appetite. Can we conclude that the percentage of children experiencing a loss of appetite from Prevnar is significantly less than for other medications?