Statistics review

via Spiegel, Schiller, Srinivasan (2009) Schaum's Probability and Statistics (third edition)

Ian Stewart

August 7, 2017

These notes represents my study of statistics during the summer of 2017, drawing on the excellent statistics primer *Schaum's Probability and Statistics (third edition)* (Spiegel, Schiller, Srinivasan (2009)) particularly for the sample problems. I'm going to skip a lot of useful details like derivations and proofs in order to focus on application rather than theory, but I recommend Wasserman's (2004) *All of Statistics* for an in-depth mathematical treatment of statistical theory. Some would say that without all the proofs, there cannot be a complete understanding of statistics. I agree, but I also think that a surface-level understanding of statistics is often sufficient to do the hypothesis testing and analysis involved in social science research. You can never learn too much, but you also have to determine for yourself how far down the rabbit hole you want to go.

The examples that I use come from social science and natural language processing contexts, because that's what I do.

1 Useful distributions

Here I outline the basic ideas of some useful probability distributions and explain their application to sample problems.

1.1 Binomial

The binomial distribution models the joint probability of X successes in n Bernoulli trials, where the probability of success is static over time. We say that p is the probability of success in a single trial, and q is the probability of failure in a single trial.

Sample problem 4.68

Let's say the probability of survival within 30 years for 30-year-olds in population P is $\frac{2}{3}$. Not sure why statisticians are so morbid, but we'll go with it for now. What is the

PDF	$P(X = x) = \binom{n}{x} p^x q^{n-x}$
μ	np
σ	$\sqrt{npq}$

Table 1: Useful facts about the binomial distribution.

probability that, out of a sample of 5 30-year-olds: (a) 5 survive, (b) more than 3 survive, (c) 2 survive, (d) at least one survives?

(a) $P(X = 5) = \binom{5}{5}(\frac{2}{3})^5(\frac{1}{3})^0 = \frac{32}{243} = 13.2\%$

(b) $P(X \geq 3) = P(X = 4) + P(X = 5) = \binom{5}{4}(\frac{2}{3})^4(\frac{1}{3})^1 + \binom{5}{5}(\frac{2}{3})^5(\frac{1}{3})^0 = 79.0\%$

(c) $P(X = 2) = \binom{5}{2}(\frac{2}{3})^5(\frac{1}{3})^0 = 16.5\%$

(d) $P(X \geq 1) = \sum_{1 \leq i \leq 5} P(X = i) = 1 - P(X = 0) = \binom{5}{0}(\frac{2}{3})^0(\frac{1}{3})^5 = 99.6\%$

Application The binomial distribution is often used to test whether an event is occurring more frequently than by chance. In the context of the sample problem, a study may find that the number of 30-year-olds in a certain subpopulation far exceeds the number expected by previous findings, by comparing the actual number of survivals with the expected survival count ($\mu = np$) generated by the binomial distribution.

1.2 Multinomial

The multinomial distribution models the probability of i separate independent random variables X_i occurring n_i times over n Bernoulli trials, such as drawing differently colored marbles out of a (large) bag. We say that p_i is the probability of success for variable X_i in a single trial, and that q_i is the probability of failure.

PDF	$P(X_i = n_i) = n!\Pi\frac{p_i^{n_i}}{n_i!}$
μ_i	np_i
σ_i	$\sqrt{np_iq_i}$

Table 2: Useful facts about the multinomial distribution.

Sample problem 4.96

Let's say the population P of a large fictional country is geographically divided such that the ratio of South (S) to North (N) to East Coast (E) to West Coast (W) is 4:3:2:1. Drawing a sample of size $n = 10$, what is the probability of drawing: (a) 4 South, 3 North, 2 East and 1 West, (b) 8 South, 2 West?

For both (a) and (b), we compute the probabilities of drawing a single member of one of the subpopulations as:

$p_S = \frac{4}{4+3+2+1} = \frac{4}{10} = 0.4, p_N = 0.3, p_E = 0.2, p_W = 10\%$
(a) $P(n_S = 4, n_N = 3, n_E = 2, n_W = 1) = 10!\frac{0.4^4 0.3^3 0.2^2 0.1^1}{4!3!2!1!} = 3.48\%$
(b) $P(n_S = 8, nW = 2) = 10!\frac{0.4^8 0.1^2}{8!2!} = 0295\%$

Application The multinomial distribution is frequently used in topic modeling situations (e.g., Mei, Shen and Zhai (2007)) because each document in a collection can be said to be drawn from a multinomial distribution of a fixed number of topics.

1.3 Hypergeometric

The hypergeometric distribution models the probability of drawing X successes in n Bernoulli trials from a population of size N, without replacement. For instance, drawing sample individuals from a population without replacement. Just like the binomial distribution, we say p is the probability of observing success in one trial, and q is the probability of observing failure.

PDF	$P(X = x) = \frac{\binom{Np}{x}\binom{Nq}{n-x}}{\binom{N}{n}}$
μ	np
σ	$\sqrt{\frac{npq(N-n)}{N-1}}$

Table 3: Useful facts about the hypergeometric distribution.

Sample problem 4.98

The population P contains 15 individuals, with 5 vegetarians and 10 omnivores. In a random sample of 8 individuals, what is the probability of drawing: (a) 4 vegetarians? (c) at least 1 vegetarian?
(a) $P(X = 4) = \frac{\binom{5}{4}\binom{10}{8-4}}{\binom{15}{8}} = \frac{70}{429} = 16.3\%$
(b) $P(X \geq 1) = 1 - P(X = 0) = 1 - \frac{\binom{5}{0}\binom{10}{8}}{\binom{15}{8}} = 1 - \frac{1}{13*11} = 99.3\%$

1.4 Geometric

The geometric distribution models the probability of undergoing X Bernoulli trials until the first success. Just like the binomial distribution, we say that p is the probability of success and q is the probability of failure.

Sample problem 4.52

PDF	$P(X = x) = pq^{x-1}$
μ	$\frac{1}{p}$
σ	$\sqrt{\frac{q}{p^2}}$

Table 4: Useful facts about the geometric distribution.

We have a text corpus of pizza toppings (each a unigram token), of which there are 6 unique types in equal proportions. Sampling repeatedly from the corpus, what is the probability of drawing “anchovies” on the fifth sample?
$P(X = anchovies) = (\frac{1}{6})(\frac{5}{6})^4 = 0.0804$

1.5 Negative binomial

The negative binomial distribution models the probability of observing X Bernoulli trials until the r^{th} success. The geometric distribution is a special case of this distribution in which $r = 1$.

PDF	$P(X = x) = \binom{x-1}{r-1} p^r q^{x-r}$
μ	$\frac{r}{p}$
σ	$\frac{\sqrt{rq}}{p}$

Table 5: Useful facts about the negative binomial distribution.

Sample problem Using the same text corpus as in the geometric distribution problem (6 toppings in equal proportions), what is the probability of drawing 20 tokens until we observe 5 “anchovies” tokens?
$p = \frac{1}{6}, r = 5, x = 20$
$P(X = 20) = \binom{19}{4} \frac{1}{6}^5 \frac{5}{6}^{15} = 3876 \frac{5^{15}}{6^{20}} = 3.24\%$

1.6 Poisson

The Poisson distribution models the probability of observing X successes in a sample occurring at a fixed rate λ in a given timespan. The Poisson distribution can approximate the binomial distribution in cases with a large n ($n \geq 50$) and a low success rate ($p < 0.1$), which means that λ can be approximated as np.

Sample problem 4.93

Over a corpus of pizza topping words, the average rate of seeing the word “mushroom” is 3 per 100,000 tokens. In a sample of 200,000 tokens, what is the probability of observing (a) zero “mushroom” tokens, (e) between 4 and 8 “mushroom” tokens?

PDF	$P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$
μ	λ
σ	$\sqrt{\lambda}$

Table 6: Useful facts about the Poisson distribution.

$p = \frac{3}{100000}, \lambda = np = 200,000(\frac{3}{100000}) = 6$
(a) $P(X = 0) = \frac{6^0 e^{-6}}{0!} = e^{-6} = 0.248\%$
(b) $P(4 \leq X \leq 8) = \sum_{4 \leq x \leq 8} \frac{\lambda^x e^{-6}}{x!} = e^{-6} \sum_{4 \leq x \leq 8} \frac{\lambda^x}{x!} = 69.6\%$

Application The Poisson distribution has been used to model the average word frequency rate of individual users online (Altmann, Pierrehumbert and Motter 2011), which makes sense if you consider each individual user to have a different but constant rate of using a particular word.

1.7 Normal

The normal distribution is the most widely used continuous probability function due in part to its ubiquity in natural phenomena. After estimating a fit to a normal distribution, it is common to convert value x to its standard normal form $Z = \frac{x-\mu}{\sigma}$, which is normally distributed with $\mu = 0$ and $\sigma = 1$.

PDF	$P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)/2\sigma^2}$
μ	λ
σ	$\sqrt{\lambda}$

Table 7: Useful facts about the normal distribution.

Sample problem 4.81
A population P uses the word “y’all” for the second-person plural, at an average rate $\mu = 61.4\%$ with standard deviation $\sigma = 0.25\%$. What is the proportion of the population that uses “y’all” (a) between 61% and 61.8% of the time? (b) exactly 61.5% of the time?
Let’s say x is the proportion of “y’all”-ers.
(a) $P(0.61 \leq x \leq 0.618) = P(\frac{0.61-0.614}{0.0025} \leq Z \leq \frac{0.618-0.614}{0.0025}) = P(-1.6 \leq Z \leq 1.6) = P(Z \geq -1.6) - P(Z \geq 1.6) = 89.04\%$
(b) $P(x = 0.615) = P(Z = \frac{0.615-0.614}{0.0025}) = P(Z = 0.4) = 36.8\%$

Applications The normal distribution shows up in a lot of places, due in part to the Central Limit Theorem which shows that the sum of a set of independent and identically distributed (IID) variables tends toward a normal distribution. Furthermore, many statistical analysis techniques such as a linear regression rely on assumptions of normality.

1.8 Chi-square

The chi-square distribution models the distribution of a sum χ^2 of the squares of independent normally distributed variables $\{\forall v X_v\}$:

$\chi^2 = \sum_v X_v^2$

The parameter v is known as the degrees of freedom, which makes sense if you think about higher v indicating more variability in the underlying X values and thus more "freedom" in the distribution of χ^2.

PDF	$P(X = x) = \frac{1}{2^{v/2}\Gamma(v/2)} x^{(v/2)-1} e^{-x/2}$
μ	v
σ	$2v$

Table 8: Useful facts about the chi-squared distribution.

Sample problem 4.111

Assume $\{\forall v X_v\}$ is a set of independent normally distributed random variables. If we look at the 99^{th} percentile of the distribution of the sum of X_v, what is the value of χ^2 that we would observe with: (a) 8 degrees of freedom? (b) 28 degrees of freedom?

Just to be clear, we are looking up chi-squared values in a table of percentile values versus degrees of freedom, such that the area under the distribution curve is at the 99^{th} percentile.

(a) $\chi^2_{0.99,8} = 20.1$

(b) $\chi^2_{0.99,28} = 48.3$

Applications The chi-squared distribution is often used in hypothesis testing to compare combinations of normal distributions, such as Chancellor, Mitra and De Choudhury (2016) in which the difference between two predictive models' fit to data ("deviance") follows a chi-squared distribution with degrees of freedom equal to the difference in the number of predictive variables.

1.9 Student's t

Student's t distribution is used to approximate a normal distribution with a known mean and unknown standard deviation (over sample $\{\forall_i x_i\}$, sample mean $\bar{X}$, sample variance S, sample size n):

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$$

where $n-1=v$ is the degrees of freedom. When $v > 30$, the t distribution approximates the normal distribution, so the t distribution is often used in cases where the sample size is small. The distribution is symmetric, so $t_{v,p} = -t_{v,1-p}$ and $P(c \leq t) = 1 - P(-c \leq t)$ for a constant c.

PDF	$P(X=x) = \frac{\Gamma(\frac{v+1}{2})}{\sqrt{v\pi}\Gamma(\frac{v}{2})}(1+\frac{x^2}{v})^{-\frac{v+1}{2}}$
μ	0
σ	$\frac{v}{v-2}$

Table 9: Useful facts about student's t distribution.

Sample problem *4.119*

If variable U is drawn from student's distribution with $v=10$, find constant c such that: (a) $P(U > c) = 0.05$, (b) $P(-c \leq U \leq c) = 0.98$, (c) $P(U \leq c) = 0.20$

(a) $c = t_{8,1-0.05} = t_{8,0.95} = 1.86$

(b)

$$P(-c \leq U \leq c) = P(c \leq U) - P(-c \leq U) = 1 - 2(P(-c \leq U)) = 0.98, P(-c \leq U) = 0.01, P(c \leq U) = 0.99 t_{10,0.}$$

(c)

$$P(U \leq c) = 0.20, P(c \leq U) = 1 - 0.20 = 0.80, t_{10,0.80} = 0.879, t_{10,0.20} = -0.879$$

Applications This distribution is often used in the comparison of two sample means whose standard deviations are assumed to be identical, such as Pavalanathan and Eisenstein (2016) who use a paired t-test to compare the distribution of emoticon use between a control and treatment population.

1.10 F

The F distribution models the distribution of the ratio of chi-squared variables V_1 and V_2, with respective degrees of freedom v_1 and v_2, such that

$$\frac{V_1/v_1}{V_2/v_2} = V \sim F_{v_1,v_2}$$

Note the use of multiple degrees of freedom v_1, v_2.

The F distribution is not symmetric but the reciprocal statistic value can be computed as follows: $F_{1-p,v_2,v_1} = \frac{1}{F_{p,v_1,v_2}}$

PDF	$P(X = x) = \frac{\Gamma(\frac{v_1+v_2}{2})}{\Gamma(\frac{v_1}{2})\Gamma(\frac{v_2}{2})} v_1^{v_1/2} v_2^{v_2/2} u^{(v_1/2)-1} (v_2 + v_1 x)^{(v_1+v_2)/2}$
μ	$\frac{v_2}{v_2-2}$ $(v_2 > 2)$
σ	$\sqrt{\frac{2v_2^2(v_1+v_2-2)}{v_1(v_2-4)(v_2-2)^2}}$ $(v_2 > 4)$

Table 10: Useful facts about the F distribution.

Sample problem 4.47 (a) For chi-squared variables V_1 with 10 degrees of freedom and V_2 with 15 degrees of freedom, compute the 95^{th} percentile of their expected ratio.

(d) For chi-squared variables V_1 with 15 degrees of freedom and V_2 with 9 degrees of freedom, compute the 1^{st} percentile of their expected ratio.

(a) $\frac{V_1/10}{V_2/15} = F_{10,15}, F_{0.95,10,15} = 2.54$

(d) $\frac{V_1/15}{V_2/9} = F_{15,9}, F_{0.01,15,9} = \frac{1}{F_{0.99,9,15}} = \frac{1}{3.89} = 0.257$

Applications The F distribution is used in the analysis of variance when a sample's standard deviation is unknown, such as Prabhakaran, Reid and Rambow (2014) which used an ANOVA test to compare the group means of different predictive features to assess which features significantly differentiated different subgroups.

2 Sampling

- To perform statistical inference on a population, it is often necessary to draw a representative *sample*, which may be drawn with or without replacement.

- Sample statistics are estimates of the population's underlying distribution. Formally, for a sample of random variables $\{X_1 ... X_n\}$, a statistic $g(X_1 ... X_n)$ approximates a parameter from the population distribution.

- We often need to estimate the distribution of sample statistics in order to draw conclusions about a population. For instance, computing the variance of the sample mean tells us how spread out the estimation of the population mean is.

- **Estimating normal distribution parameters**

- For sample mean $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$, the expected value is $\mu_{\bar{X}} = \mu$. The variance of the sampling distribution is $\sigma_{\bar{X}} = \frac{\sigma^2}{n}$ with replacement, and $\sigma_{\bar{X}} = \frac{\sigma^2}{n}\frac{N-n}{N-1}$.

- For sample variance $S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$, the expected value with replacement is $\mu_{S^2} = \frac{n-1}{n}\sigma^2$, which indicates that S^2 is a biased estimator because its expected value is not equal to the population parameter (as compared to $\hat{S}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$). The expected value without replacement is $\mu_{S^2} = (\frac{N}{N-1})(\frac{n-1}{n})\sigma^2$. The variance of the variance is $\sigma_{S^2} = \frac{\sigma}{\sqrt{2n}}$.

- We already know that the standardized variable $Z = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is normally distributed. But if we do not know population variance σ^2, we can estimate using statistic $T = \frac{\bar{X}-\mu}{\hat{S}/\sqrt{n}} = \frac{\bar{X}-\mu}{S/\sqrt{n-1}}$, which follows student's t distribution with $n-1$ degrees of freedom.

- For two independent populations, we can estimate the expected value and variance of the difference between population samples' sample means $\bar{X_1}, \bar{X_2}$ are $\mu_{\bar{X}_1-\bar{X}_2} = \mu_1 - \mu_2$ and $\sigma_{\bar{X}_1-\bar{X}_2} = \sqrt{\sigma^2_{\bar{X}_1} + \sigma^2_{\bar{X}_2}} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$. The expected value and variance of the difference between population samples' sample standard deviations S_1, S_2 are $\mu_{S_1-S_2} = \mu_{S_1} - \mu_{S_2}$ and $\sigma_{S_1-S_2} = \sqrt{\sigma^2_{S_1} - \sigma^2_{S_2}}$. The standardized variable is normally distributed if n_1 and n_2 are over 30 and can be computed as $Z = \frac{(\bar{X}_1-\bar{X}_2)-(\mu_1-\mu_2)}{\sqrt{\sigma_1^2/n_1+\sigma_2^2/n_2}}$.

- For two independent populations, we can estimate the ratio of the population variances $\frac{S_1^2}{S_2^2}$ with test statistic $F = \frac{n_1 S_1^2/(n_1-1)\sigma_1^2}{n_2 S_2^2/(n_2-1)\sigma_2^2} = \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$.

- **Estimating binomial distribution parameters**

- For sample proportion of success P, we can estimate the expected value and variance as $\mu_P = p$ and $\sigma_P = \sqrt{\frac{pq}{n}} = \sqrt{\frac{p(1-p)}{n}}$, derived from population parameter p and identities $\mu = p, \sigma = \sqrt{p(1-p)}$.

- For two independent populations, we can estimate the expected value and variance of the difference between populations' proportions of success P_1, P_2 as $\mu_{P_1-P_2} = \mu_{P_1} + \mu_{P_2} = p_1 - p_2$ and $\sigma_{P_1-P_2} = \sqrt{\sigma^2_{P_1} + \sigma^2_{P_2}} = \sqrt{\frac{p_1q_1}{n_1} + \frac{p_2q_2}{n_2}}$.

Sample problems

5.54

A population X has mean weekly income of \$800 with standard deviation \$60. What is the probability that a random sample of 16 people has a mean daily income (a) between \$790 and \$810? (c) greater than \$820?

(a) $P(790 \leq \bar{X} \leq 810) = P(\frac{790-800}{60/\sqrt{16}} \leq Z \leq \frac{810-800}{60/\sqrt{16}}) = P(-\frac{2}{3} \leq Z \leq \frac{2}{3}) = P(Z \leq \frac{2}{3}) - P(Z \leq -\frac{2}{3}) = 0.4972$

(c) $P(\bar{X} \geq 820) = P(Z \geq \frac{820-800}{60/\sqrt{16}}) = P(Z \geq \frac{4}{3}) = 1 - P(Z \leq \frac{4}{3}) = 0.0912$

5.60

A corpus contains 80 tokens, 60% nouns and 40% verbs. Out of 50 samples of 20 tokens (with replacement), how many samples will have (a) equal numbers of nouns and verbs? (c) 8 nouns and 12 verbs?

Let's call noun probability $p = 0.60$ and verb probability $q = 0.40$.

$\mu_{\bar{X}} = 0.60(20) = 12, \sigma_{\bar{X}} = \sqrt{npq} = \sqrt{20(0.60)(0.40)} = 2.19089$

We use the normal approximation of the binomial distribution, $Z_{\bar{X}} = \frac{\bar{X}-\mu}{\sigma/\sqrt{n}}$.

(a) $P(\bar{X} = 10) = P(Z = \frac{10-12}{2.19089}) = P(Z = -0.912871) = 0.263$, therefore the expected number of samples with equal counts of nouns and verbs is $50(0.263) = 18$.

(c) $P(\bar{X} = 8) = P(Z = \frac{8-12}{2.19089}) = P(Z = -1.8257) = 0.075$, therefore the expected number of samples with 8 nouns and 12 verbs is $50(0.075) = 4$.

These are the wrong answers, they should be (a) 6 and (c) 2. Why??

5.68

A large corpus contains 65% punctuation. Find the probability that two random samples, each comprising 200 tokens, have a greater than 10% difference in proportions of punctuation tokens.

$n = 200, \mu_{P_1-P_2} = 0.65 - 0.65 = 0.0, \sigma_{P_1-P_2} = \sqrt{\frac{0.65(0.35)}{200} + \frac{0.65(0.35)}{200}} = 0.047697.$

$P(\mu_{P_1} - \mu_{P_2} \geq 10\%) = P(Z \geq \frac{0.10-0.}{0.047697}) = P(Z \geq 2.09657) = 0.01801.$

Answer should be 0.0410.

5.75

The fundamental frequency of the same vowel extracted from a population of (helium-addicted) speakers follows a normal distribution of 2000 Hz and a standard deviation of 60 Hz. If 10 speakers are selected at random, what is the probability that the sample standard deviation will (a) not exceed 50 Hz, (b) lie between 50 and 70 Hz?

$n = 10, \mu_S = \sigma = 60, \sigma_S = \frac{\sigma}{\sqrt{2n}} = \frac{60}{\sqrt{2(10)}} = 13.4164.$

(a) $P(S \leq 50) = P(Z \leq \frac{50-60}{13.4164}) = P(\frac{-10}{13.4164}) = 0.2280.$

(b) $P(50 \leq S \leq 70) = P(\frac{50-60}{13.4164} \leq Z \leq \frac{70-60}{13.4164}) = P(-0.74536 \leq Z \leq 0.74536) = 0.5439$

Answers should be (a) 0.36 (b) 0.49.

5.80

Two large Twitter corpora, T and W, contain tweets whose length is normally distributed. The standard deviation of tweet length in T is 40 characters, and the standard deviation of tweet length in W is 50 characters. We take a sample of 8 tweets from T and 16 tweets from W. What is the probability that the variance of the sample from T is more than (a) 2, (b) 1.2 times the variance from the sample from W?

$\sigma_T = 40, \sigma_W = 50, n_T = 8, n_W = 16$

We can estimate the probability of these variance ratios with the F statistic:

$$F = \frac{n_T S_T^2/((n_T - 1)\sigma_T^2)}{n_W S_W^2/((n_W - 1)\sigma_W^2)} F = \frac{8/((8-1)40^2)}{16/((16-1)50^2)}(\frac{S_T^2}{S_W^2}) = \frac{1/1400}{4/9375}(\frac{S_T^2}{S_W^2}) = \frac{375}{224}(\frac{S_T^2}{S_W^2})$$

(a) $F = \frac{375}{224}(2) = 3.3482 \sim F_{7,15}, P(F_{7,15} \geq 3.3482) = 0.02$
(b) $F = \frac{375}{224}(1.2) = 2.00892 \sim F_{7,15}, P(F_{7,15} \geq 2.00892) = 0.12$

5.130

In a subcommunity on Twitter, the hashtag #notallmen is normally distributed among the community members with mean frequency 72 and standard deviation 8. (a) Find the minimum frequency of the top 20% of members. (b) Find the probability that in a random sample of 100 students, the minimum frequency among the top 20% will be less than 76.

(a) High-level: we want to find the frequency that splits the lower 80% of members from the upper 20%, which is the same thing as finding the point in the normal distribution that separates the lower 80% of cumulative probability from the remaining 20%.

$P(Z' \geq Z) = 80\%$, where $Z' = \frac{X-\mu}{\sigma} = \frac{X-72}{8}$. Using the inverse CDF for the normal distribution where $P(Z' \geq Z) = 80\%$, we find $Z' = 0.842$, so we can solve for X as follows: $X = Z'(\sigma) + \mu = (0.842)(8) + 72 = 78.736 \approx 79$.

(b) Slightly different problem! We want to find $P(M_{20} \leq 76)$ where M_{20} is the minimum frequency among the top 20% of members. If $\bar{X}$ is the sample mean, $M_{20} = \bar{X} + m$, where m is the amount needed to reach M_{20} from $\bar{X}$. We can standardize this as $P(\bar{X} + m \leq 76) = P(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} + m' \leq \frac{76-\mu}{\sigma/\sqrt{n}}) = P(Z + m' \leq \frac{4}{8/10})$, therefore $P(Z \leq 5 - m')$. Since m' is the difference between the 50% and 80% mark, we know that $m' = 0.842$, therefore $P(Z \leq 5 - 0.842) = P(Z \leq 4.158) = (1 - 0.99998) = 0.00002$.

Answer should be 0.0090...why?

3 Estimation

- A statistic is an *unbiased* estimator if the expected value of the statistic is equal to the parameter (e.g., $E(\bar{X}) = \mu$.
- A statistic is more *efficient* if it has a smaller variance (e.g., $V(\mu) < V(med)$).
- Establishing *confidence intervals* for a statistic S lets us know how likely it is for us to find the actual population parameters within those intervals. For instance, for the statistic $\bar{X}$, we can be 68.27% confident of finding μ in the interval $\bar{X} - \sigma_{\bar{X}} < \bar{X} < \bar{X} + \sigma_{\bar{X}}$. 95% confident $= S \pm 1.96\sigma_S$, 99% confident $= S \pm 2.58\sigma_S$.
- z_c is equal to the Z-value corresponding to the desired confidence percentile c
- For a normal distribution, the confidence interval for the population mean is $\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}}$, assuming a large sample or sampling with replacement. For sampling without replacement, for population size N the confidence interval is $\bar{X} \pm z_c \frac{\sigma}{\sqrt{n}} \sqrt{\frac{N-n}{N-1}}$.
- For a normal distribution with a small sample size ($n < 30$), we have to use a t-distribution approximation: $-t_c < \frac{(\bar{X}-\mu)\sqrt{n}}{\hat{S}} < t_c, \bar{X} \pm t_c \frac{\hat{S}}{\sqrt{n}}$.

- To approximate the proportion p of a binomial distribution, we use $P \pm z_c\sqrt{\frac{p(1-p)}{n}}$ for sampling with replacement, and $P \pm z_c\sqrt{\frac{pq}{n}\frac{N-n}{N-1}}$ for sampling without replacement.
- To find the confidence interval on the sum of two independent sample statistics S_1 and S_2, just use the sum of their variances: $S_1 + S_2 \pm z_c\sqrt{\sigma_{S_1^2} + \sigma^2_{S_2}}$. For example, the confidence interval on the sum of two sample means is $\bar{X}_1 + \bar{X}_2 \pm z_c\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$.
- To compute the confidence interval for standard deviation, we use the fact that $\frac{(n-1)\hat{S}^2}{\sigma^2}$ follows a chi-squared distribution with $n-1$ degrees of freedom. Specifically, using confidence level c and splitting it between the two tails of the chi-squared distribution (i.e. using $\chi_{c/2}$ and $\chi_{1-c/2}$), we have $\chi^2_{c/2} \leq \frac{(n-1)\hat{S}^2}{\sigma^2} \leq \chi^2_{1-c/2}$, $\frac{\hat{S}\sqrt{n-1}}{\chi_{1-c/2}} \leq \sigma \leq \frac{\hat{S}\sqrt{n-1}}{\chi_{c/2}}$.
- To compute the confidence interval on the ratio of variances. we use the fact that $\frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} \sim F_{n_1-1,n_2-1}$. Specifically, using confidence level c split between the two tails of the F distribution, we have $F_{c/2} \leq \frac{\hat{S}_1^2/\sigma_1^2}{\hat{S}_2^2/\sigma_2^2} \leq F_{1-c/2}$, or $\frac{1}{F_{1-c/2}}\frac{\hat{S}_1^2}{\hat{S}_2^2} \leq \frac{\sigma_1^2}{\sigma_2^2} \leq \frac{1}{F_{c/2}}\frac{\hat{S}_1^2}{\hat{S}_2^2}$.
- To extract a point estimate of a population parameter rather than confidence-interval estimate, we often use the *maximum likelihood estimate.* If a population has a parameter θ to be estimated by some statistic, the probability of observing datum x is computed with the density function $f(x, \theta)$. The likelihood of observing n independent data $\{x_1 ... x_n\}$ is $L = \prod_{1 \leq i \leq n} f(x_i, \theta)$, or in log-likelihood is $log(L) = \sum_{1 \leq i \leq n} log(f(x_i, \theta)$. We solve for θ by taking the derivative of the likelihood, setting to 0, and solving for θ, i.e. $\sum_{1 \leq i \leq n} \frac{1}{f(x_i,\theta)}\frac{\delta f(x_i\theta)}{\delta\theta} = 0$.

Sample problems

6.31

A population of English speakers uses the word “fetch” with a (sample) mean annual frequency of 1200 and a standard deviation of 100.

(a) Estimate the mean and standard deviation of the population assuming the samples are taken with size (i) 30, (ii) 50, and (iii) 100. (b) What is the relationship between sample standard deviations and estimates of population standard deviations for different sample sizes?

(a) We estimate the mean for all sample sizes as $\bar{X} = \frac{\sum_{1 \leq i \leq n} X_i}{n} = \frac{1200(n)}{n} = 1200$. We estimate the standard deviation $\hat{s} = \sqrt{\frac{n}{n-1}}s$, which is (i) $\sqrt{\frac{30}{29}}100 = 101.71$, (ii) $\sqrt{\frac{50}{49}}100 = 101.02$, (iii) $\sqrt{\frac{100}{99}}100 = 100.50$.

(b) We can conclude that the estimates of population standard deviation become more accurate with larger sample sizes.

6.35

If the standard deviation of a population of humans' lifetimes is estimated as 100 years, how large must the sample be to have (a) 95%, (c) 99% confidence that the error in the estimated mean lifetime will not exceed 20 years?

We know that the estimation error for sample mean is $z_c\frac{\sigma}{\sqrt{n}} \leq 20$, so we need to substitute the appropriate z_c values given the confidence levels and solve for $n \geq (z_c\frac{\sigma}{20})^2$.

(a) $z_{0.95} = 1.96, n \geq (1.96\frac{100}{20})^2 = 96.04$, therefore $n = 97$.

(c) $z_{0.99} = 2.58, n \geq (2.58\frac{100}{20})^2 = 166.41$, therefore $n = 167$.

6.39

Five measurements of an individual's pronunciation of the phoneme /a/ were recorded as 0.28, 0.30, 0.27, 0.33, 0.31 seconds in duration. Find (a) 95%, (b) 99% confidence limits for the actual mean reaction time.

This counts as a small sample ($n = 5$), so we need to use the approximation $\bar{X} \pm t_c\frac{\hat{S}}{\sqrt{n}}$. $\bar{X} = 0.298, \hat{S} = 0.21354, \frac{\hat{S}}{\sqrt{n}} = \frac{0.21354}{\sqrt{5}} = 0.009550$.

(a) 95% split over two tails yields $t_{0.975} = 2.78, CI = 0.298 \pm 2.78(0.00955) = 2.98 \pm 0.027$.

(c) 99% split over two tails yields $t_{0.995} = 4.60, CI = 2.98 \pm 4.60(0.00955) = 2.98 \pm 0.044$.

6.40

A Basque corpus contains absolutive and ergative nouns in an unknown proportion. A random sample of 60 verbs selected with replacement from the corpus showed that 70% of nouns were absolutive. Find the (a) 95%, (c) 99.73% confidence limits for the true proportion of absolutive nouns.

$P = 0.70, CI = P \pm z_c\sqrt{\frac{p(1-p)}{n}} = 0.70 \pm z_c\sqrt{\frac{0.70(0.30)}{60}} = 0.70 \pm z_c(0.05916)$.

(a) $z_{0.95} = 1.96, CI = 0.70 \pm 1.96(0.05916) = 0.70 \pm 0.1160$.

(c) $z_{0.99} = 2.58, CI = 0.70 \pm 2.58(0.05916) = 0.70 \pm 0.1526$.

6.44

A sample of 200 verbs from a corpus showed that 15 were defective, while a sample of 100 verbs from another corpus showed that 12 were defective. Find (b) 99%, (c) 99.73% confidence limits for the difference in proportions of defective verbs from the two corpora.

$P_1 = \frac{15}{200} = 0.075, P_2 = \frac{12}{100} = 0.12, CI = P_1 - P_2 \pm z_c\sqrt{\frac{P_1(1-P_1)}{n_1} + \frac{P_2(1-P_2)}{n_2}} = P_1 - P_2 \pm z_c(0.03746)$

(b) $z_{0.99} = 2.58, CI = -0.045 \pm 2.58(0.03746) = -0.45 \pm 0.09665$.

(c) $z_{0.9973} = 3.00, CI = -0.45 \pm 3.00(0.03746) = -0.45 \pm 0.1124$.

6.48

The standard deviation of the frequency of 10 verbs in a corpus is 120. Find (a) 95%, (b) 99% confidence limits for the standard deviation of all verbs in the corpus.

$CI = S \pm z_c\frac{\sigma}{\sqrt{2n}} = z_c\frac{120}{\sqrt{20}} = z_c(26.83)$.

(a) $z_{0.95} = 1.96, CI = [120 - 1.96(26.83), 120 + 1.96(26.83)] = [67.41, 172.6]$.

(b) $z_{0.99} = 2.58, CI = [120 - 2.58(26.83), 120 + 2.58(26.83)] = [50.77, 189.2]$.
(a) [87.0, 230.9], (b) [78.1, 288.5]