Stat 381: Applied Statistical Methods I **Spring 2015**

March 9-20

Instructor: Brian Powers

16.1 Random Sampling

Definition 16.1. A random variable X which can take any real number value from a **population**.

Definition 16.2. A **sample** is a subset of a population A sampling procedure which produces inferences that consistently over or underestimate some characteristic of the population are said to be **biased**. A **random sample** is chosen so that the observations are independent and at random. A random sample of size n is

$$X_1, X_2, \ldots, X_n$$

with numerical values $x_1, x_2, \ldots, x_n$. Random variables in a random sample are said to be **independent and identically distributed (iid)**.

16.2 Some Important Statistics

An estimate of a population parameter is given the hat as an identifier. For example, the estimate of a population proportion p is $\hat{p}$, read "p hat".

Definition 16.3. any function of the random variables from a random sample is a **statistics** Recall the following sample statistics of the **location**

Definition 16.4. The **sample mean** $\bar{X}$ is

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Definition 16.5. The **sample median** is

$$\tilde{x} = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd,} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

Definition 16.6. The **sample mode** is the value of the sample which occurs most often.

Definition 16.7. The **sample variance** S^2 is

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Definition 16.8. The **sample standard deviation** is $S = \sqrt{S^2}$.

Definition 16.9. The **sample range** is $X_{max} - X_{min}$.

16.3 Sampling Distributions

Given a random sample of size n from a population with mean μ and variance σ^2, what is the mean and variance of $\bar{X}$?

$$E(\bar{X}) = E\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n}\left(E(X_1) + \cdots + E(X_n)\right) = \frac{1}{n} nE(X_1) = \mu$$

$$Var(\bar{X}) = Var\left(\frac{1}{n}(X_1 + \cdots + X_n)\right) = \frac{1}{n^2}\left(Var(X_1) + \cdots + Var(X_n)\right) = \frac{1}{n^2} nVar(X_1) = \frac{\sigma^2}{n}$$

Definition 16.10. The probability distribution of a statistic is a **sampling distribution**.

16.3.1 Properties of Some Distributions

Theorem 16.11. *If $X_1, \ldots, X_n$ are independent, and $X_i \sim N(\mu_i, \sigma_i^2)$, then*

$$\sum_{i=1}^{n} X_i \sim N\left(\sum \mu_i, \sum \sigma_i^2\right)$$

Theorem 16.12. *If $X_1, \ldots, X_n$ are independent, and $X_i \sim Gamma(\alpha_i, \beta)$, then*

$$\sum_{i=1}^{n} X_i \sim Gamma\left(\sum \alpha_i, \beta\right)$$

Corollary 16.13. *If $X_1, \ldots, X_n \sim Exp(\beta)$ are iid, then*

$$\sum_{i=1}^{n} X_i \sim Gamma\left(n, \beta\right)$$

16.3.2 The Central Limit Theorem

Theorem 16.14. ***Central Limit Theorem:*** *If $\bar{X}$ is the sample mean of a sample of size n, from a population with mean μ and variance σ^2, then*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution as $n \to \infty$.

Example 16.15. At a pencil company the machines are supposed to produce pencils of average length 20cm. The pencils have a standard deviation $\sigma^2 = .2$ cm. A random sample of 100 pencils is found to have a mean length of 20.13 cm. Is there reason to believe that the machines are not calibrated correctly?

Example 16.16. A punk band's songs are on average 1:44 with a standard deviation of 10 seconds. If you make a random mix of 40 of their songs, what is the probability it will last longer than 75 minutes?

16.3.3 Difference of Sample Means

Corollary 16.17. *If independent random samples of sizes n_1 and n_2 are drawn from two populations with respective means μ_1, μ_2 and variances σ_1^2, σ_2^2, the difference of the sample means $\bar{X}_1 - \bar{X}_2$ is approximately normal (moreso as $n_i \to \infty$) with*

$$\mu_{\bar{X}_1 - \bar{X}_2} = \mu_1 - \mu_2, \quad \sigma^2_{\bar{X}_1 - \bar{X}_2} = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

so

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

16.3.4 Sampling Distribution of S^2

Recall that

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X - \bar{X})^2.$$

By Adding and subtracting $\bar{X}$, we can write

$$\begin{aligned}
\sum_{i=1}^{n}(X_i - \mu)^2 &= \sum_{i=1}^{n}\left[(X_i - \bar{X}) + (\bar{X} - \mu)\right]^2 \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{n}(\bar{X} - \mu)^2 + 2(\bar{X} - \mu)\sum_{i=1}^{n}(X_i - \bar{X}) \\
&= \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu)^2
\end{aligned}$$

We substitute $(n-1)S^2 = \sum(X_i - \bar{X})$ and divide both sides by σ^2.

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \mu)^2 = \frac{(n-1)S^2}{\sigma^2} + \frac{(\bar{X} - \mu)^2}{\sigma^2/n}$$

Left hand side follows a Chi-Squared distribution with n degrees of freedom. The second term on the right is Z^2 which is Chi-Squared with 1 degree of freedom. It takes a little more theory than this course contains, but we get the following conclusion:

Theorem 16.18. *Given $X_1, \ldots, X_n$ iid from a Normal population with variance σ^2,*

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

follows a Chi-Squared distribution with $n-1$ degrees of freedom.

Example 16.19. Car batteries have a lifetime that is normally distributed, with a supposed standard deviation of 1 year. If 5 batteries are sampled with lifetimes of 1.9, 2.4, 3, 3.5 and 4.2 years, should we suspect that the standard deviation has changed?

16.4 t-Distribution

Often the population variance is unknown, so it is natural to use S^2 as an estimate. So we use

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

But for small samples, the value of S may vary quite a bit from sample to sample. This statistic follows what is known as a t-distribution.

Theorem 16.20. *If $X_1, \ldots, X_n$ are iid $N(\mu, \sigma^2)$, then*

$$T = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a t-distribution with $v = n - 1$ degrees of freedom.

Even when the population is not normal, if it is approximately normal (bell shaped, symmetric) then the distribution will be approximately a t-distribution.