# STAT 704 --- Preliminaries: Basic Inference

## Basic Definitions

**Random variable: A function that maps an outcome from some random phenomenon to a real number.**
**• A r.v. measures the result of a random phenomenon.**

**Example 1: The weekly income of a randomly selected USC student.**
**Example 2: The number of accidents in a month at a busy stretch of highway.**

**• Every r.v. (say, *Y*) has a probability density function (pdf) *f*(*y*) associated with it.**
**For a discrete r.v. *Y*,**

**For a continuous r.v. *Y*, and for any numbers $a < b$,**

**Expected Value: The expected value of a r.v. is the mean of its probability distribution.**

**For a discrete r.v. *Y*,**

**For a continuous r.v. *Y*,**

**Note: If *a* and *c* are constants,**

**<u>Variance</u>: The <u>variance</u> of a r.v. measures the "spread" of its probability distribution.**

**var(*Y*) =**

**=**

**Equivalently,**

**Note: If *a* and *c* are constants,**

**Note: The standard deviation of a r.v. *Y* is**

**Example: Suppose *Y* (the high temperature in Celsius of a random September day in Seattle) has expected value 20 and variance 10. Let *W* = the high temperature in Fahrenheit. Then**

**<u>Covariance</u>: For two r.v.'s *Y* and *Z*, the <u>covariance</u> of *Y* and *Z* is**

**• If *Y* and *Z* have _________ covariance, then small values of *Y* tend to correspond to ________ values of *Z* (and large values of *Y* to ________ values of *Z*).**

**Example:**

**• If $Y$ and $Z$ have _________ covariance, then small values of $Y$ tend to correspond to ________ values of $Z$ (and large values of $Y$ to ________ values of $Z$).**

**Example:**

**Note: If $a_1$, $c_1$, $a_2$, and $c_2$ are constants,**

**Note:**

**• The correlation coefficient between $Y$ and $Z$ is similar, but is scaled to be between –1 and 1:**

**If corr($Y$, $Z$) = 0, then we say**

**<u>Independent Random Variables</u>: Informally, two r.v.'s $Y$ and $Z$ are independent if knowing the value of one r.v. does not affect the probability distribution for the other r.v.**

**Note: If $Y$ and $Z$ are independent, then**

**• A covariance of zero does <u>not</u> imply independence in general, but…**

**• If $Y$ and $Z$ are <u>normal</u> r.v.'s, then**

## Linear Combinations of Random Variables

• **Suppose** $Y_1, Y_2, \ldots, Y_n$ **are r.v.'s and** $a_1, a_2, \ldots, a_n$ **are constants.**

**Then**

**Important Example: Suppose** $Y_1, Y_2, \ldots, Y_n$ **are independent r.v.'s, each with expected value μ and variance σ². Then consider the sample mean** $\overline{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ **:**

**Central Limit Theorem**: When we take a large sample of size ***n*** from a population with mean μ and variance $\sigma^2$, and ***n*** is "reasonably large", then $\bar{Y}$ has an approximately normal distribution with mean μ and variance $\dfrac{\sigma^2}{n}$.

**The Normal Distribution**: A r.v. ***Y*** having a normal distribution has the pdf:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu)^2}$$, for any $-\infty < y < \infty$

• **The two parameters of a normal distribution are its mean μ and its variance $\sigma^2$.**

• **If *Y* ~ N(μ, $\sigma^2$), then the standardized r.v.**

**Note: If *a* and *c* are constants and *Y* is a normal r.v., then**

**Note: If** $Y_1, Y_2, \ldots, Y_n$ **are independent normal r.v.'s and** $a_1, a_2, \ldots, a_n$ **are constants, then** $a_1Y_1 + \cdots + a_nY_n$

**Example: Suppose** $Y_1, Y_2, \ldots, Y_n$ **are a random sample from a normally distributed population with mean μ and variance $\sigma^2$. Then**

## Other Related Distributions

**Chi-square**: **If** $Z_1, \ldots, Z_\nu$ **are independent N(0, 1) r.v.'s, then**

**<u>t-distribution</u>: If $Z$ and $X$ are independent r.v.'s and**
**$Z$ ~ N(0, 1) and $X$ ~ $\chi^2_\nu$, then**

**<u>F-distribution</u>: Suppose $X_1$ and $X_2$ are independent r.v.'s and**
**$X_1$ ~ $\chi^2_{\nu_1}$, $X_2$ ~ $\chi^2_{\nu_2}$. Then**

**<u>Note</u>: The square of a**

**<u>Proof</u>:**

## <u>A Model for a Single Sample</u>

**• Suppose we have a random sample $Y_1, Y_2, \ldots, Y_n$ of observations from a normal distribution with unknown mean $\mu$ and unknown variance $\sigma^2$.**
**• We can model this as:**

**• Often we wish to perform inference (confidence interval or hypothesis test) about the unknown population mean $\mu$.**

**Let**

**<u>Fact</u>:**

**<u>Heuristic "Proof"</u>:**

**<u>Fact</u>:** $\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ **has a distribution.**

**Therefore, look at:**

**We see that** $\dfrac{\bar{Y}-\mu}{s/\sqrt{n}}$ **has a distribution.**

**(Note that $\bar{Y}$ and $s^2$ are independent when we sample from a normal distribution.)**

**So, under model (*),**

**<u>CI Example</u> (Summer temperatures data):**

**Interpretation:**

**(See R example on course web page.)**

## <u>Hypothesis Testing</u>

**• We may also perform a t-test to determine whether $\mu$ may equal some specified value, say $\mu_0$.**

**• We decide whether to reject a <u>null hypothesis</u> ($H_0$) about $\mu$ on the basis of our sample evidence (as measured by our <u>test statistic</u>).**

**• Let**

**Three types of test:**

**• Note that under $H_0$ (if $\mu$ really is $\mu_0$), then $t^*$ has a**

**• If the $t^*$ that we observe is highly unusual (relative to the Distribution), we will reject $H_0$ and conclude $H_a$.**

**• Let $\alpha$ = the <u>significance level</u> = maximum allowable probability of rejecting $H_0$ when $H_0$ is true.**

## Rejection rules

**Two-sided:**

**One-sided ($H_a$: “<”):**

**One-sided ($H_a$: “>”):**

**P-value approach: We can also measure the evidence against $H_0$ using a P-value, which is the probability of observing a test statistic as extreme or more extreme than the test statistic value that we did observe, if $H_0$ were true.**

**• A small P-value indicates strong evidence against $H_0$.**

**Rule:**

**• The calculation of the P-value depends on the alternative hypothesis:**

**$H_a$: “$\neq$”**

**$H_a$: “<”**

**$H_a$: “>”**

**Example: We wish to test whether the true mean high temperature is greater than 75 degrees, using $\alpha = 0.01$.**

**<u>Conclusion</u>:**

## <u>Connection between CIs and Two-sided tests</u>

**<u>Fact</u>: An $\alpha$-level two-sided test rejects $H_0$: $\mu = \mu_0$ <u>if and only if</u> $\mu_0$ falls <u>outside</u> a $(1 - \alpha)100\%$ CI about $\mu$.**

**Previous example: At $\alpha = 0.10$, would we reject $H_0$: $\mu = 73$ and conclude $H_a$: $\mu \neq 73$?**

**At $\alpha = 0.10$, would we reject $H_0$: $\mu = 80$ and conclude $H_a$: $\mu \neq 80$?**

**At $\alpha = 0.05$, would we reject $H_0$: $\mu = 80$ and conclude $H_a$: $\mu \neq 80$?**

## <u>Paired Data</u>

**• When we have two <u>paired samples</u> (when each observation in one sample can be naturally paired with an observation in the other sample), we can typically use our one-sample methods to conduct inference on the <u>mean difference</u>.**

**<u>Example</u>: 7 pairs of mice were injected with a cancer cell. Mice within each pair came from the same "litter" and were therefore similar biologically. For each pair, one mouse was given an experimental drug and the other mouse was untreated. After a specific time, the tumors were weighed.**

**Let $Y_{1j}$ =**

**Let $Y_{2j}$ =**

**• Since these samples are paired, we take the differences**

**If the differences follow a normal distribution, then we have the model:**

**• To test whether the treatment results in a lower mean tumor weight, we can test:**

**Example:**

# <u>Two Independent Samples</u>

**• Assume we now have two independent (not paired!) samples from two normal populations. Label them sample 1 and sample 2.**

**Model:**

**<u>Note</u>: Both populations have the same variance, $\sigma^2$.**
**<u>Note</u>: The two sample sizes ($n_1$ and $n_2$) may be different.**

**An estimator of the variance $\sigma^2$ is the**

**Then**

**• Our parameter of interest is the difference in the two population means, $\mu_1 - \mu_2$.**

**A $(1 - \alpha)100\%$ CI for $\mu_1 - \mu_2$ is**

**• Often we wish to test whether the two populations have the same mean.**
**• We test**
**against one of the following alternatives, using the test statistic:**

**$H_a$:** **Rejection Rule:** **P-value**

## Case of Unequal Variances

• **What if it is not reasonable to assume the two populations have the same variance? Suppose**

• **Use**

• **The standard deviation part of the test statistic is now**

• **Our test statistic under $H_0$ has an approximate t-distribution with d.f. given by an approximation formula (Satterthwaite's formula or Welch's formula).**

**Model:**

• **We can formally test** $H_0 : \sigma_1^2 = \sigma_2^2$ **using an F-test, but in practice graphical methods, e.g., box plots, are often employed.**

• **R and SAS perform the two-sample t-test, with options for the equal-variance case and the unequal-variance case.**

**Example: Testing pollution levels: 10 pollution measurements were taken upstream of a chemical plant, and 15 measurements were taken downstream. Do the mean pollution levels differ ($\alpha = 0.05$)?**

**<u>Note</u>: Recall our t-procedures require that the data come from a <u>normal</u> population.**
**• Fortunately, the t-procedures are robust: They work approximately correctly if the population distribution is "close" to normal.**
**• Also, if our sample size is large, we can use the t procedures even if our data are not normal (related to CLT).**
**• If the sample size is small, we should perform some check of the normality assumption before using t-procedures.**

## Normal Q-Q plots

**• To use the t-procedures (test and CI), the assumption that the data come from a normal population must be reasonable.**
**• Could check with a**

**(Verify distribution is**

**• More precise plot: Normal Q-Q plot.**
**• Plots quantiles of data against suitably chosen standard normal percentiles.**
**• If Q-Q plot resembles a straight line, then the normal assumption is reasonable.**

**• Possible violations:**

## <u>Nonparametric Tests</u>

**• If the data do not come from a normal population (and if the sample size is not large), we cannot use the t-test.**
**• Must use nonparametric ("distribution-free") methods.**

### Sign Test

**• For the sign test, we assume the data come from a continuous distribution.**
**Model:**

**• We test**

**Test statistic is**

**Under $H_0$, $B^*$ follows a**

**• Reject $H_0$ if $B^*$ is an "unusual" value relative to this distribution.**

• **Alternative could be**

**Example: (Eye relief data)**

**Wilcoxon Rank Sum Test (also known as Mann-Whitney Test)**

• **This is a test comparing the medians of two independent samples from continuous populations.**

• **We assume the two population distributions are identical except for a possible shift (if**

**Model:**

• **We test**

**Method: Rank the <u>combined</u> sample**

• **The "rank sum statistic"** ***W*** **is the sum of the ranks of the second-sample values in the combined sample.**

**• If $W$ is very large, this is evidence that**

**• If $W$ is very small, this is evidence that**

**Example (Dental measurements):**

**• Wilcoxon rank-sum test can also test whether one population is <u>stochastically larger</u> than another.**

## <u>Wilcoxon Signed-Rank Test</u>

**• This assumes the data come from a continuous, symmetric distribution.**
**• Again, we test**

**• Test statistic uses**

**• The <u>signed rank</u> for observation $i$ is**

**• The "signed rank statistic" is**

**• If $W^+$ is very large, this is evidence that**

**• If $W^+$ is very small, this is evidence that**

**• Both the sign test and the signed-rank test can be used with paired data (e.g., we could test whether the median difference is zero).**

**Example (Weather station data):**

**• The sign test and signed-rank test are more flexible than the t-test (require less strict assumptions), but the t-test has more power when the data truly have a normal distribution.**