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*, $f(y) = P(Y = y)$ for ~~all~~ any value y.

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

$$P(a < Y < b) = \int_a^b f(y)\,dy$$

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

For a discrete r.v. *Y*, the expected value is

$$E(Y) = \sum_{\text{all } y} y\, f(y)$$

For a continuous r.v. *Y*,

$$E(Y) = \int y\, f(y)\,dy$$

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

$$E[a + cY] = a + c\,E(Y)$$

In particular,

$$E(a) = a$$

$$E(cY) = c\,E(Y)$$

$$E(Y + a) = E(Y) + a$$

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

var(*Y*) = $E\{[Y - E(Y)]^2\}$

= the "expected squared deviation" of Y from its mean.

Equivalently,

$$\mathrm{var}(Y) = E(Y^2) - [E(Y)]^2$$

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

$$\mathrm{var}[a + cY] = c^2\,\mathrm{var}(Y)$$

In particular, $\mathrm{var}(a) = 0$

$$\mathrm{var}(cY) = c^2\,\mathrm{var}(Y)$$

$$\mathrm{var}(Y + a) = \mathrm{var}(Y)$$

Note: The standard deviation of a r.v. *Y* is $\sqrt{\mathrm{var}(Y)}$

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

$$E(W) = E\left(\frac{9}{5}Y + 32\right) = \frac{9}{5}E(Y) + 32 = \frac{9}{5}(20) + 32 = 68$$

$$\mathrm{var}(W) = \mathrm{var}\left(\frac{9}{5}Y + 32\right) = \left(\frac{9}{5}\right)^2 \mathrm{var}(Y) = 3.24(10) = 32.4$$

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

$$\mathrm{cov}(Y, Z) = E\{[Y - E(Y)][Z - E(Z)]\}$$

Also, $\mathrm{cov}(Y, Z) = E(YZ) - E(Y)E(Z)$

- **If *Y* and *Z* have** <u>positive</u> **covariance, then small values of *Y* tend to correspond to** <u>small</u> **values of *Z* (and large values of *Y* to** <u>large</u> **values of *Z*).**

Example: Height & Weight of People

Work Experience & Salary

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

Example: Altitude + Temperature of Mtn. Sites
Size and Gas mileage of Cars

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

$$\text{cov}[a_1 + c_1 Y, a_2 + c_2 Z] = c_1 c_2 \text{cov}[Y, Z]$$

Note: By definition, $\text{cov}(Y,Y) = \text{var}(Y)$

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

$$\text{corr}(Y,Z) = \frac{\text{cov}(Y,Z)}{\sqrt{\text{var}(Y)\,\text{var}(Z)}}$$

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

Y and Z are uncorrelated.

(Y and Z are not linearly associated)

<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 Y and Z are uncorrelated ($\text{cov}(Y,Z) = 0$)

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

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

$$\text{cov}(Y,Z) = 0 \iff Y, Z \text{ independent.}$$

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 $E\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} a_i E(Y_i)$

i.e., $E[a_1 Y_1 + \cdots + a_n Y_n] = a_1 E(Y_1) + \cdots + a_n E(Y_n)$

Also, $\text{var}\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j \, \text{cov}(Y_i, Y_j)$

For two r.v.'s: $E(Y_1 + Y_2) = E(Y_1) + E(Y_2)$

$\text{var}(Y_1 + Y_2) = \text{var}(Y_1) + \text{var}(Y_2) + 2\,\text{cov}(Y_1, Y_2)$

$\text{var}(Y_1 - Y_2) = \text{var}(Y_1) + \text{var}(Y_2) - 2\,\text{cov}(Y_1, Y_2)$

Note: If $Y_1, \ldots, Y_n$ are independent, then

$\text{var}\left[\sum_{i=1}^{n} a_i Y_i\right] = \sum_{i=1}^{n} a_i^2 \, \text{var}(Y_i)$.

Also, If $Y_1, \ldots, Y_n$ are all independent, then

$$\text{cov}\left[\sum_{i=1}^{n} a_i Y_i, \sum_{i=1}^{n} c_i Y_i\right] = \sum_{i=1}^{n} a_i c_i \, \text{var}(Y_i)$$

Important Example: Suppose $Y_1, Y_2, \ldots, Y_n$ are independent r.v.'s, each with expected value μ and variance σ^2. Then consider the sample mean $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ **:**

$E(\bar{Y}) = E\left[\frac{1}{n} Y_1 + \cdots + \frac{1}{n} Y_n\right] = \frac{1}{n} E(Y_1) + \cdots + \frac{1}{n} E(Y_n)$

$= \frac{1}{n}\mu + \cdots + \frac{1}{n}\mu = n\left(\frac{1}{n}\mu\right) = \mu$

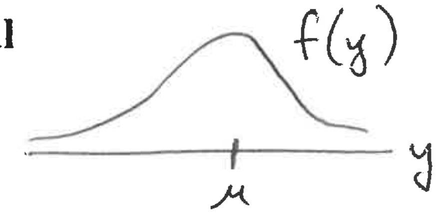
and $\text{var}(\bar{Y}) = \text{var}\left[\frac{1}{n} Y_1 + \cdots + \frac{1}{n} Y_n\right]$

$= \frac{1}{n^2}\text{var}(Y_1) + \cdots + \frac{1}{n^2}\text{var}(Y_n) = \frac{1}{n^2}\sigma^2 + \cdots + \frac{1}{n^2}\sigma^2$

$= n\left(\frac{1}{n^2}\sigma^2\right) = \frac{\sigma^2}{n}$

Central Limit Theorem: ~~**With**~~ When **we take a large sample of size n from a population with mean μ and variance σ^2, and n is "reasonably large", then $\overline{Y}$ 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:**



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

- **The two parameters of a normal distribution are its mean μ and its variance σ^2.**
- **If $Y \sim N(\mu, \sigma^2)$, then the standardized r.v.** $Z = \dfrac{Y-\mu}{\sigma}$

mean ← variance

has a "standard normal" distribution, $N(0,1)$.

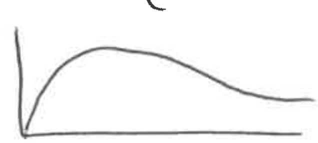
Note: If a and c are constants and Y is a normal r.v., then
$a + cY$ also has a normal distribution, with mean $a + cE(Y)$ and variance $c^2 \, var(Y)$.

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 $a_1Y_1 + \cdots + a_nY_n$ also has a normal distribution, with mean $\sum_{i=1}^{n} a_i E(Y_i)$ and variance $\sum_{i=1}^{n} a_i^2 \, var(Y_i)$.

Example: Suppose $Y_1, Y_2, \ldots, Y_n$ are a random sample from a normally distributed population with mean μ and variance σ^2. Then $\overline{Y}$ has a normal distribution with mean μ and variance σ^2/n.

Other Related Distributions

Chi-square: **If $Z_1, \ldots, Z_\nu$ are independent N(0, 1) r.v.'s, then**
$Z_1^2 + \cdots + Z_\nu^2$ has a $\chi^2_{(\nu)}$ distribution. "chi-square with ν degrees of freedom". $\chi^2_{(\nu)}$ pdf:



- The mean of a $\chi^2_{(\nu)}$ r.v. is ν.
- The variance of a $\chi^2_{(\nu)}$ r.v. is 2ν.

t-distribution: **If Z and X are independent r.v.'s and $Z \sim N(0, 1)$ and $X \sim \chi^2_\nu$, then** $\dfrac{Z}{\sqrt{X/\nu}}$ has a $t_{(\nu)}$ distribution (t with ν d.f.)

- The mean of a $t_{(\nu)}$ r.v. is 0. (if $\nu > 1$)
- The variance of a $t_{(\nu)}$ r.v. is $\dfrac{\nu}{\nu-2}$ (if $\nu > 2$)

F-distribution: **Suppose X_1 and X_2 are independent r.v.'s and** ~~**$Z \sim N(0, 1)$ and**~~ **$X_1 \sim \chi^2_{\nu_1}$, $X_2 \sim \chi^2_{\nu_2}$. Then** $\dfrac{X_1/\nu_1}{X_2/\nu_2}$ has an F_{ν_1, ν_2} distribution. Two parameters: "numerator d.f." and "denominator d.f."

Note: **The square of a** $t_{(\nu)}$ r.v. is an $F_{(1, \nu)}$ r.v.

Proof: $$[t_{(\nu)}]^2 = \left[\frac{Z}{\sqrt{\chi^2_\nu/\nu}}\right]^2 = \frac{Z^2}{\chi^2_\nu/\nu} = \frac{\chi^2_1/1}{\chi^2_\nu/\nu} = F_{1,\nu}$$

A Model for a Single Sample

- **Suppose we have a random sample $Y_1, Y_2, \ldots, Y_n$ of observations from a normal distribution with unknown mean μ and unknown variance σ^2.**
- **We can model this as:** $Y_j = \mu + \varepsilon_j, \quad j = 1, \ldots, n \qquad (*)$

  where $\varepsilon_j \sim N(0, \sigma^2)$

- **Often we wish to perform inference (confidence interval or hypothesis test) about the unknown population mean μ.**

Let $S^2 = \dfrac{\sum_{i=1}^{n} (Y_i - \bar{Y})^2}{n-1}$ be the sample variance.

and $S = \sqrt{S^2}$ be the sample standard deviation.

Fact: $\dfrac{(n-1)s^2}{\sigma^2}$ has a $\chi^2_{(n-1)}$ distribution.

Heuristic "Proof": Note: $\sum_{i=1}^{n}\left(\dfrac{Y_i-\mu}{\sigma}\right)^2 = \sum_{i=1}^{n} Z_i^2$ has a $\chi^2_{(n)}$ distribution. And $\dfrac{(n-1)s^2}{\sigma^2} = \dfrac{n-1}{\sigma^2}\sum_{i=1}^{n}\dfrac{1}{n-1}(Y_i-\bar{Y})^2$ $= \sum_{i=1}^{n}\left(\dfrac{Y_i-\bar{Y}}{\sigma}\right)^2$. This has a $\chi^2_{(n-1)}$ distribution since we lose 1 degree of freedom by estimating μ with $\bar{Y}$.

Fact: $\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}$ **has a** $N(0,1)$ **distribution.**

$N(0,1)$

Therefore, look at:

$$\frac{\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\sqrt{\dfrac{(n-1)s^2}{\sigma^2}\Big/(n-1)}} = \frac{\dfrac{\bar{Y}-\mu}{\sigma/\sqrt{n}}}{\dfrac{s}{\sigma}} = \frac{\bar{Y}-\mu}{s/\sqrt{n}}$$

$\sqrt{\dfrac{\chi^2_{(n-1)}}{n-1}}$

We see that $\dfrac{\bar{Y}-\mu}{s/\sqrt{n}}$ **has a** $t_{(n-1)}$ **distribution.**

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

So, under model (*),

$$P\left(-t_{(1-\alpha/2,\,n-1)} < \frac{\bar{Y}-\mu}{s/\sqrt{n}} < t_{(1-\alpha/2,\,n-1)}\right) = 1-\alpha$$

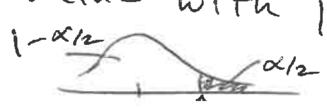
$$\Rightarrow P\left(-t_{(1-\alpha/2,\,n-1)}\frac{s}{\sqrt{n}} < \bar{Y}-\mu < t_{(1-\alpha/2,\,n-1)}\left(\frac{s}{\sqrt{n}}\right)\right) = 1-\alpha$$

$$\Rightarrow P\left(\bar{Y}-t_{(1-\alpha/2,\,n-1)}\frac{s}{\sqrt{n}} < \mu < \bar{Y}+t_{(1-\alpha/2,\,n-1)}\frac{s}{\sqrt{n}}\right) = 1-\alpha$$

So a $(1-\alpha)100\%$ CI for μ under model (*) is:

$$\boxed{\bar{Y} \pm t_{(1-\alpha/2,\,n-1)}\frac{s}{\sqrt{n}}}$$

where $t_{(1-\alpha/2,\,n-1)}$ is the value with $1-\alpha/2$ area to the left in the $t_{(n-1)}$ distn:



$1-\alpha = .90$

CI Example (Summer temperatures data):

$\bar{Y} = 77.667, \quad s = 8.872, \quad n = 30$

90% CI: $t_{(.95, 29)} = 1.699$ (Table B.2)

$77.667 \pm (1.699)\left(\frac{8.872}{\sqrt{30}}\right) \Rightarrow (74.91, 80.42)$

Interpretation: With 90% confidence, the true mean high temperature is between 74.91 and 80.42 degrees.

(See R example on course web page.)

Hypothesis Testing

- **We may also perform a t-test to determine whether μ may equal some specified value, say μ_0.**
- **We decide whether to reject a null hypothesis (H_0) about μ on the basis of our sample evidence (as measured by our test statistic).**
- **Let** $t^* = \dfrac{\bar{Y} - \mu_0}{s/\sqrt{n}}$ be our test statistic.

Three types of test:

Two-sided	(One-sided: "<")	(One-sided: ">")
$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$	$H_0: \mu = \mu_0$
$H_a: \mu \neq \mu_0$	$H_a: \mu < \mu_0$	$H_a: \mu > \mu_0$

- **Note that under H_0 (if μ really is μ_0), then t^* has a** $t_{(n-1)}$ distribution
- **If the t^* that we observe is highly unusual (relative to the** $t_{(n-1)}$ **Distribution), we will reject H_0 and conclude H_a.**
- **Let α = the significance level = maximum allowable probability of rejecting H_0 when H_0 is true.**

Rejection rules

Two-sided: If $|t^*| > t_{(1-\alpha/2,\, n-1)}$ then reject Ho, otherwise fail to reject Ho.

One-sided (H_a: "<"): If $t^* < t_{(\alpha,\, n-1)} = -t_{(1-\alpha,\, n-1)}$ then reject Ho, otherwise fail to reject Ho.

One-sided (H_a: "~~<~~>"): If $t^* > t_{(1-\alpha,\, n-1)}$ then reject Ho, otherwise fail to reject Ho.

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:

If P-value $< \alpha \Rightarrow$ reject Ho. otherwise fail to reject Ho.

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

H_a: "≠" $\Rightarrow$ P-value = 2 times the area outside our t^* in the $t_{(n-1)}$ distribution.

H_a: "<" $\Rightarrow$ P-value = Area to the left of our t^* in the $t_{(n-1)}$ distribution.

H_a: ">" $\Rightarrow$ P-value = Area to the right of our t^* in the $t_{(n-1)}$ distribution.

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

$H_0: \mu = 75$ vs. $H_a: \mu > 75$

$$t^* = \frac{77.667 - 75}{8.872/\sqrt{30}} = 1.646$$ Note $t_{(.99,\, 29)} = 2.462$

$1.646 \not> 2.462 \Rightarrow$ fail to reject Ho.

From R, P-value $= .055 > .01 = \alpha$.

Conclusion: We fail to reject Ho. At $\alpha = .01$, we cannot say the true mean temperature is greater than $75°$.

Connection between CIs and Two-sided tests

Fact: An α-level two-sided test rejects H_0: $\mu = \mu_0$ if and only if μ_0 falls outside a $(1 - \alpha)100\%$ CI about μ.

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

At $\alpha = 0.10$, would we reject H_0: $\mu = 80$ and conclude H_a: $\mu \neq 80$? No

At $\alpha = 0.05$, would we reject H_0: $\mu = 80$ and conclude H_a: $\mu \neq 80$? No

Paired Data

- **When we have two paired samples (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 mean difference.**

Example: 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} = tumor weight of mouse j receiving "control"

Let Y_{2j} = tumor weight for mouse j receiving "treatment"

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

$$D_j = Y_{1j} - Y_{2j}$$

If the differences follow a normal distribution, then we have the model: $D_j = \mu_D + \varepsilon_j, \quad j = 1, \ldots, n$

where $\varepsilon_j \sim N(0, \sigma^2)$ and $\mu_D = \mu_1 - \mu_2$

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

$$H_0: \mu_D = 0 \quad \text{vs.} \quad H_a: \mu_D > 0$$

Example: For this test on the mice data, the P-value is 0.0008 (and $t^* = 5.39$). At $\alpha = 0.05$, we can reject H_0 and conclude the true mean difference is greater than 0 (the treatment produces a significantly lower mean tumor weight).

Verify: $t^* > t_{0.95, 6}$

- From R, a 95% CI for the true mean difference μ_D is: $(0.273, 0.725)$. The mean tumor weight for untreated mice is between 0.27 grams higher and 0.73 grams higher than the mean tumor weight for treated mice.

Two Independent Samples

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

Model: $Y_{ij} = \mu_i + \varepsilon_{ij} \quad i = 1, 2 \quad j = 1, 2, \ldots, n_i$

where $\varepsilon_{ij} \sim N(0, \sigma^2)$

Note: Both populations have the same variance, σ^2.
Note: The two sample sizes (n_1 and n_2) may be different.

An estimator of the variance σ^2 is the "pooled sample variance"

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}$$

Then $\dfrac{(\bar{Y}_1 - \bar{Y}_2) - (\mu_1 - \mu_2)}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}}$ has a $t_{(n_1+n_2-2)}$ distn.

• 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

$$(\bar{Y}_1 - \bar{Y}_2) \pm t_{(1-\alpha/2,\, n_1+n_2-2)} \sqrt{\frac{S_p^2}{n_1} + \frac{S_p^2}{n_2}}$$

• Often we wish to test whether the two populations have the same mean.

• We test $H_0: \mu_1 = \mu_2$ (equivalently: $H_0: \mu_1 - \mu_2 = 0$)
against one of the following alternatives, using the test statistic:

$$t^* = \frac{(\bar{Y}_1 - \bar{Y}_2)}{\sqrt{\dfrac{S_p^2}{n_1} + \dfrac{S_p^2}{n_2}}}$$

H_a:	Rejection Rule:	P-value
$\mu_1 \neq \mu_2$	Rej H_0 if $\lvert t^* \rvert > t_{(1-\alpha/2,\, n_1+n_2-2)}$	2 times area outside t^* in $t_{(n_1+n_2-2)}$ distn.
$\mu_1 < \mu_2$	Rej H_0 if $t^* < -t_{(1-\alpha,\, n_1+n_2-2)}$	Area to left of t^* in $t_{(n_1+n_2-2)}$ distn
$\mu_1 > \mu_2$	Rej. H_0 if $t^* > t_{(1-\alpha,\, n_1+n_2-2)}$	Area to right of t^* in $t_{(n_1+n_2-2)}$ distn.

Case of Unequal Variances

- **What if it is not reasonable to assume the two populations have the same variance? Suppose** $\sigma_1^2 \neq \sigma_2^2$
- **Use** s_1^2 and s_2^2 to estimate σ_1^2 and σ_2^2.

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

$$\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

- **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: $Y_{ij} = \mu_i + \varepsilon_{ij} \qquad i = 1, 2 \qquad j = 1, \ldots, n_i$

$\varepsilon_{ij} \sim N(0, \sigma_i^2)$

- **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.**

$H_0: \mu_1 = \mu_2$ vs. $H_a: \mu_1 \neq \mu_2$

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 ($\alpha = 0.05$)? Assuming $\sigma_1^2 = \sigma_2^2$ (based on boxplot evidence), Our $t^* = -4.64$ and p-value $= .0001$. Using $\alpha = .05$, we reject H_0 and conclude the true mean upstream pollution level differs from the true mean downstream pollution level.

<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.**

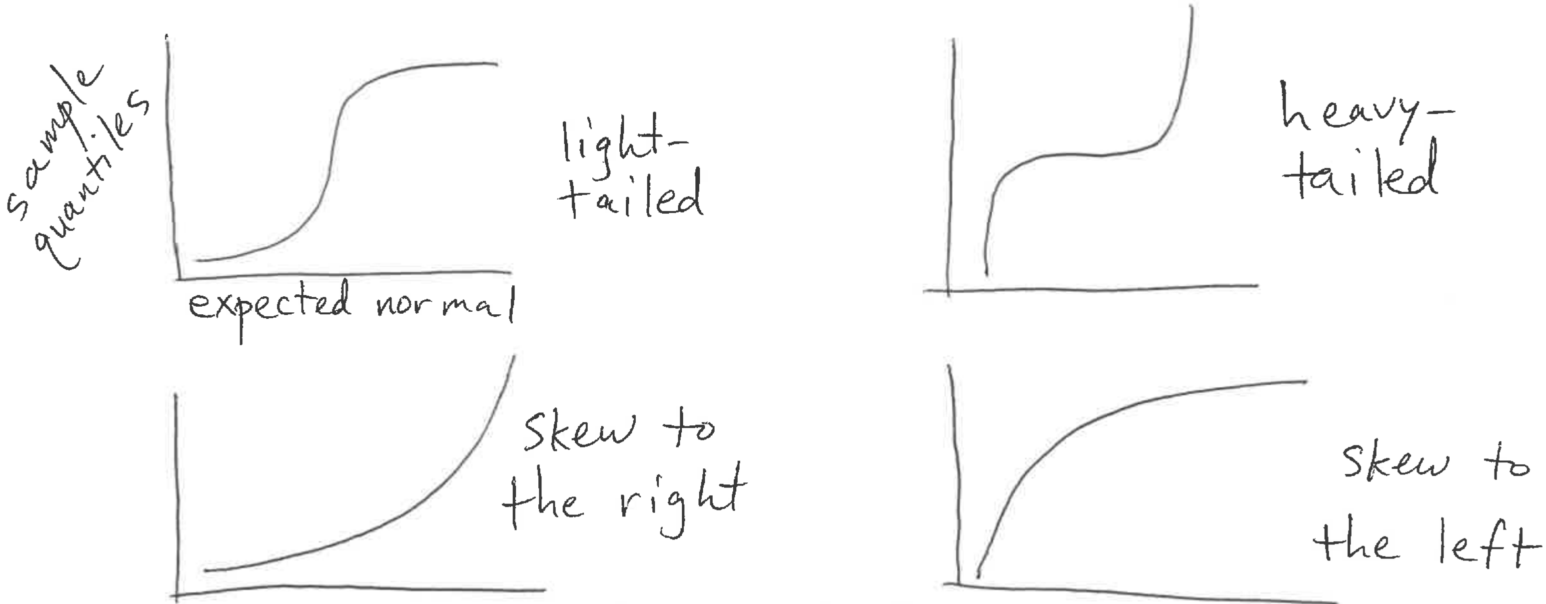
Normal Q-Q plots

- **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** histogram or stemplot or boxplot
  ↑ best of these

(Verify distribution is symmetric with no severe outliers

- **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: $Y_j = \eta + \varepsilon_j$, $j=1,\ldots,n$

η = population median and ε_j follow a continuous distn.

• **We test** $H_0: \eta = \eta_0$ where η_0 is a specific number.

Test statistic is B^* = number of $y_j > \eta_0$

(for how many values is $y_j - \eta_0$ positive?)

Under H_0, B^* follows a binomial$(n, 0.5)$

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

• **Alternative could be** $H_a: \eta < \eta_0$

or $H_a: \eta > \eta_0$

or $H_a: \eta \neq \eta_0$

Example: (Eye relief data)

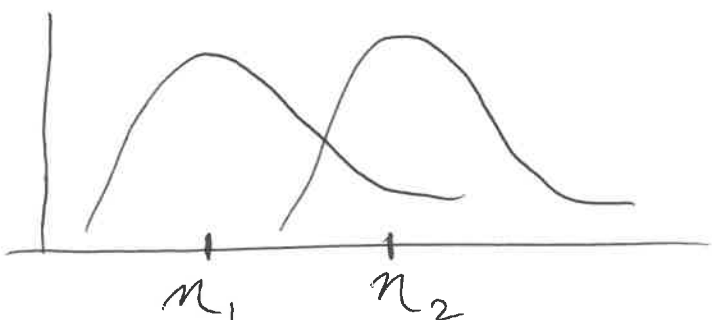
$H_0: \eta = 5$ vs. $H_a: \eta \neq 5$ at $\alpha = .05$.

From R, p-value $= 0.041$, so reject H_0. Conclude true median time-to-relief is not 5 minutes.

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** $\eta_1 \neq \eta_2$**)**



Model:

$$Y_{ij} = \eta_i + \varepsilon_{ij} \qquad i=1,2, \quad j=1,\dots,n_i$$

ε_{ij} independent from continuous population with median zero

• **We test** $H_0: \eta_1 = \eta_2$ against

$H_a: \eta_1 < \eta_2$, $H_a: \eta_1 > \eta_2$, or $H_a: \eta_1 \neq \eta_2$

Method: Rank the <u>combined</u> sample $\{y_{11}, \dots, y_{1n_1}, y_{21}, \dots, y_{2n_2}\}$

• **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** $\eta_1 < \eta_2$
(the "big" ranks tend to come from the second sample)

• **If W is very small, this is evidence that** $\eta_1 > \eta_2$

Example (Dental measurements):

$W = 150.5$, P-value $= 0.001$ from R. boys=1 girls=2

$H_0: \eta_1 = \eta_2$ vs. $H_a: \eta_1 > \eta_2$. Reject H_0, conclude median for boys is greater than Median for girls.

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

Wilcoxon Signed-Rank Test

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

• **Again, we test** $H_0: \eta = \eta_0$

• **Test statistic uses** $R_1^{(abs)}, \ldots, R_n^{(abs)}$, the ranks of $\{|y_1 - \eta_0|, \ldots, |y_n - \eta_0|\}$

• **The signed rank for observation i is**

$$R_i^+ = \begin{cases} R_i^{(abs)} & \text{if } y_i - \eta_0 \text{ is positive} \\ 0 & \text{otherwise} \end{cases}$$

• **The "signed rank statistic" is**

$$W^+ = \sum_{i=1}^{n} R_i^+$$

• **If W^+ is very large, this is evidence that** the ~~[illegible]~~ true median is greater than η_0.

• **If W^+ is very small, this is evidence that** the true median is less than η_0.

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

Example (Weather station data): Let $D = (\text{weather} - \text{pasture})$

Testing $H_0: \eta_D = 0$ vs. $H_a: \eta_D \neq 0$.

From R, P-value = .013. So at $\alpha = .05$, we reject H_0 and conclude the median difference is not zero.

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